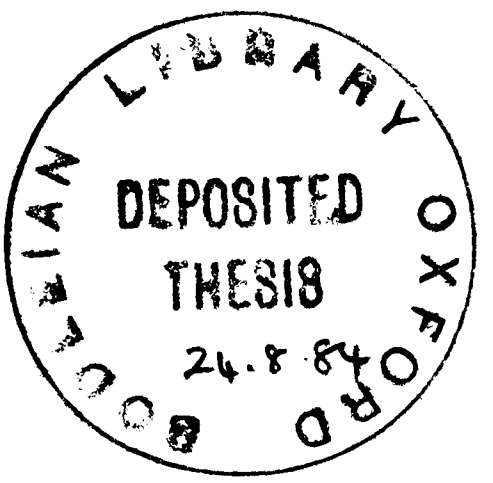
REPRESENTATION AND RATIONALITY:

FOUNDATIONS OF COGNITIVE SCIENCE

David Kirsh

Wadham College



Thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, Trinity Term, 1983.

University of Oxford

Abstract

In this essay I consider the foundations of a particular approach to cognitive science. There is a belief, among some, that cognitive research ought to proceed in two steps. In the first step, cognitive systems ought to be interpreted as rational beings, endowed with knowledge of their environments and motivated by goals and desires. Questions about the modularity of knowledge and about the sort of knowledge that is necessary or sufficient for a given competence are addressed at this stage. In the second step, inferences are to be made about the design of control systems that might instantiate these kinds of knowledge states, goals and rationality. Studies at the knowledge level are meant to serve as guidelines in the search for mechanisms.

Problems arise as soon as one asks for the justification of this view. For instance, one often hears that knowledge is attributed by a process of interpretation that is subjective; it depends on the scientist having a 'manual' for interpreting behaviour that has no foundation in fact. Knowledge is essentially observer relative; it designates nothing intrinsic in a system. In Chapter Three I argue that this position is false. By introducing the notion of 'robustness' as the touchstone of realism I suggest that knowledge states are potentially as robust as any in science.

Moreover realism about knowledge does not entail accepting what Fodor has called the Language of Thought hypothesis. We can reason about knowledge states in abstraction from the various ways knowledge can be implemented in a system. The language of thought is just one of many ways that knowledge can be used by a system. Hence there is no simple way to move from an account of what a system knows to how it uses or has access to that knowledge. In Chapter One I argue that the step from 'knowledge theories' to 'process theories' is more complicated than language of thought theorists suppose.

In Chapter Two I discuss the basic methodology of research at the knowledge level. Any well-defined task imposes severe constraints on the way it can be accomplished. The discovery of these constraints and the consequences that flow from them is perhaps the central job of knowledge level research.

I conclude the thesis with two chapters on the limitations of knowledge level research. Given that the more structured and rigid a task environment is, the more determinate the knowledge that is necessary or sufficient for task competence, we would expect that tasks and environments which are more open-ended, less closed to intervention from outside interference, would not submit to knowledge level research. Relying on a distinction between peripheral and central cognitive faculties, I question the prospects for knowledge level research of central faculties. Unlike the problem of vision or muscular co-ordination, the problem of deliberation is radically open-ended. Too many factors might become relevant to bound the class of task knowledge that might become vital.

Acknowledgments

I am fortunate to have had the help and support of friends and several members of the faculty. Prof. Sir A.J. Ayer has been constant in his supervision and friendship since I arrived in Oxford many years ago and has given me the benefit of his seasoned perspective on countless occasions. He also read this thesis and provided helpful comments. The late Gareth Evans subjected many of the ideas presented here to relentless questioning over weekly intervals during a two year period. David Wiggins offered valuable aid when the focus of the thesis was on levels of analysis. And more lately, Mark Sainsbury has helped me immeasurably in putting together a thesis with a beginning, middle and ending. He has a rare combination of patience, intelligence and compassion.

To Richard Young I owe thanks for introducing me to A.I. To Adrian Cussins I owe the pleasure of several years discussion, and gratitude for his many critical comments on this work. Rob Holte and Alison Sajovic commented on Chapters One to Four and gave me valuable assistance in integrating my ideas into a single framework. Peter Nicholas and especially Joy de Beyer were of help when I was writing an earlier version.

My strongest and deepest debt, however, is to my parents and my sisters who have offered me support and encouragement at every stage.

CONTENTS

Introduction 1

Chapter 1: INTERLEVEL RELATIONS 16

- Some Historical Background 17
- Supervenience is Too Weak 22
- Design Constraints: An Example 25
- The Formality Condition 29
- The Space of Possible Mechanisms 32
- Example 1: Procedural vs Declarative Represention 34
- Example 2: Amgiguity and Ellipsis 38
- Learning 42
- Production Systems 45
- Parallel Processing 47
- Conclusion 48

Chapter 2: CENTRAL CONCEPTS 49

- Knowledge Attribution 59
- The Task Environment 64
- Natural Tasks 70
- The Task of the Semantic Processor 75
- Marr: Knowledge of Necessary Conditions 81
- Knowledge of Sufficient Conditions 94

Chapter 3 : IS KNOWLEDGE REAL? 106

- Cognitive Ethology 108
- How Does Knowledge Constrain? Some Analogies 120
- Representational Realism 130
- The Intellectualist Fallacy 135
- A Rejoinder 137
- Another Argument 140
- When Is Tacit Knowledge Robust? 146
- Acquiring Tacit Knowledge of Technical Concepts 155

Chapter 4 : ARE THERE TWO ORDERS IN NATURE ? 163

Peripheral vs Central Cognitive Science 166

Mechanistic Explanation 169

The Rational Order 183

Rational Management of Beliefs 189

A Wittgensteinian Problem for Belief Management 199

Chapter 5 : EMPATHY 206

How Might Reasoning Be Non-Computational? 209

Explaining Action 217

Problems With Rational Decision Theory 226

The Consequence Function 228

The Openness of Ends 233

Our Biology Shows Through 236

Chapter 6 : CONCLUSION 241

Bibliography

INTRODUCTION

Cognitive science is a new field. According to Kuhn, a science only reaches maturity by passing through an immature phase where there is overt disagreement over the fundamentals. Lacking a consensus on both the legitimate problems of the subject and the proper methods of research, scientists in immature disciplines appeal to general philosophical principles in arguing for their research strategies. Decisions concerning the relevance of certain phenomena, the propriety of importing techniques and methods developed in different fields, the desirability of providing explanations of the same type as those found in other sciences, these and more must be justified. And in the early phase of a science there is no guarantee that all parties will make the same decision.

It is widely accepted that the ultimate goal of cognitive science is to produce a detailed theory of the structure and processes of the human cognitive system. Since the human mind is the most complex control system we know, the techniques and methods of research required to investigate it may differ from those found in other fields. And indeed this is so. Modern cognitive science rests on the premise that mental processes can be described and explained in the language of computation. Studies therefore focus on:

(1) Specification: the description of exactly what the system is doing;

2) Representation: the analysis of how information is encoded or structured by the system; and

3) Control: the analysis of the sequence of rules or computations the system activates during behaviour.

The rationale for this division flows from the very conception of a computational device. Turing, the father of computation theory, based his precise definition of computation on an analysis of what a human being actually does when he computes.[1] He noted that such a person follows a set of rules in a completely mechanical manner. Whether he is carrying out long division, performing an algebraic manipulation, or doing a calculus problem, we see him write symbols on, say, a piece of paper, and change his behaviour as he notes various specific symbols appearing as results of computational steps. Apparently, underpinning the performance of mathematical tasks there is:

a) a set of rules and dispositions to apply rules; and,

b) a manner of systematically representing the problem or initial state, the intermediate steps along the way to its solution, and the solution itself, the output state/state.

Let us call the set of rules plus dispositions, the '<u>control structure</u>' of the calculator, and its manner of representing states its <u>notational system</u>. Turing saw the standard problems in computation theory to consist in finding for given mathematical problems, either the control structure and notational system of a machine able to solve

---

1 See Martin Davis, "What is a computation?", in Lynn Steen (ed) <u>Mathematics Today</u>, Vintage Books, 1980. pp 241-268. See also Joseph Weizenbaum, <u>Computer Power and Human Reason</u>. (San Francisco: W.H. Freeman, 1976), chap 2. And Marvin Minsky, <u>Computation: Finite and Infinite Machines</u>. (London: Prentice Hall, 1967), chap 5.

the problem efficiently, or a proof that no such system exists. Specification entered the equation because the goal of designing a control structure plus notational system is to produce a system that efficiently performs what it is supposed to do. To design a system of the appropriate sort, the designer must know the sorts of behaviour the system is to be capable of producing -- its behavioural repetoire; the sorts of circumstances it can react to -- its input sensitivity; and the relationship that is supposed to exist between the two -- its input-output function. If we did not know what a system is meant to do, how could we prove that on the basis of its inputs and control structure it could accomplish its task?

The position I shall be defending in this essay is that:

1) the specifications in cognitive science are not purely behavioural, but epistemic: they ought to describe cognitive faculties as organized bodies of knowledge.

2) the knowledge states specifying cognitive faculties are real, intrinsic states of mind-like systems; and

3) specifications in terms of knowledge cast light on the likely mechanisms responsible for behavioural performance.

Essentially the view is a modification of Chomsky's.[2] Chomsky argued that to understand linguistic competence we must give up the idea that language is a system of habits, conditioned responses, or dispositions to verbal behaviour. Language is the manifestation of a system of knowledge, specifically knowledge of grammar; and the correct

---

2 Chomsky has been expounding and defending his ideas on the nature of linguistic theory and cognitive science since the publication of Syntactic Structures (The Hague: Mouton), 1957. My chief sources have been Aspects of the Theory of Syntax (Cambridge: MIT Press), 1965; Language and Mind, enlarged edition (New york: Harcourt and Brace Jovanovich) 1972; and Rules and Representations (Oxford: Basil Blackwell), 1980.

approach to linguistics is through a study of linguistic knowledge. Accordingly, the proper subject of linguistic inquiry is not behaviour per se, but speakers' organization of their linguistic knowledge.

Moreover, linguistic knowledge is not a fictional attribute; it is a real state of speaker/hearers. In Chomsky's metaphysics, linguistic theory, no less than physics or chemistry, aims at truth. It may be a fact that truth is more than we can knowingly have. As Van Fraassen[3] has argued, the epistemic commitment involved in accepting a scientific theory does not have to be belief that it is true but only the weaker belief that it is empirically adequate. Yet, for anyone with a more realistic turning, it is natural to see science as aiming at true descriptions of unobservable processes. This is Chomsky's view. He has argued, on numerous occasions, that linguistic theories offer potentially true descriptions of unobservable states of mind. These states 'enter into' the processes of language use. More importantly, they shape and structure those processes.

The degree to which this move represents a radical departure from tradition is not always appreciated. Psychological research, since the time of Hull and Tolman, has aimed, more or less, at providing models of the control mechanisms regulating behaviour. When fully explicit, these models provide process explanations of behaviour. They indicate, in reasonably precise terms, the sequence of processes occurring in organisms which is responsible for their behaviour. Obviously with such emphasis on models that are behaviourally adequate, specifications

---

3 Van Fraassen, Bas C, The Scientific Image. (Oxford: Clarendon Press) 1980. See, esp. Chap 1.

of behavioural dispositions are crucial. In fact, they are more than crucial. All control mechanisms, whatever their complexity, are defined in terms of functionally individuated parts. To functionally individuate a part, one must state its relation to input, output and other parts. Thus, insofar as psychologists aim at process models, and all process models are functional models, specifications must of necessity be of input-output relations: in other words, specifications of behavioural dispositions.

By leaving the behavioural domain, Chomsky, in effect, was giving up the standard methodology of functionalism. He was offering a new approach to psychology and cognitive science. At times he minimised this distinction, arguing that his theories were just different sorts of functional theories, or preludes to functional theories. Yet in this he was wrong. If it is true, as he often maintained, that it is not necessary that knowledge be linked with a definite set of behavioural dispositions, then it cannot be necessary that knowledge be identical with a functionally individuated part of the system. On the contrary, knowledge states would not be functional states, they would be states of another sort, cognitive states, which the agent has the power to 'call on' in the course of speaking, thinking and reasoning. Knowledge states are more abstract than functional states; in principle, the same knowledge state can be called on in different ways, and used by different functional parts or different functional systems. Yet if knowledge states are not functional states, and they are not ways of precisely describing behaviour, what are they? Since they cannot serve as ingredients in process explanations what is their

status in a sober science?

This thesis may be regarded as an attempt at answering that question. I start from the assumption that attributions of knowledge are made from a different level of analysis -- a different analytical perspective -- than familiar analyses of function and functional states. I then explore how attribution of knowledge states can be justified, tested for realism and related to functional state. Once knowledge is regarded as belonging to a different level of analysis -- the Knowledge Level -- to use Alan Newell's phrase[4] -- we cannot assume that theories of knowledge will be easily related to theories at other levels.

Chomsky has been uncharacteristically quiet about the nature of this relation. The account which, I believe, comes closest to his intention holds that knowledge states are really functional states under an interpretation. Assuming that the mind operates in much the same way as a serial computer, we can identify functional states of the mind with sentences in a notational system. Each syntactic unit of a notational system, plays a definite role in a program. It mediates input and output in a certain way. In certain programs each syntactic unit can also be interpreted as performing an operation that is meaningful in some external domain. For example, the command 2 add 2, in most programs, can be interpreted as a command to add the number 2 to itself. In the machine there are only tokens of numerals and tokens of addition signs. But we can, if we wish, interpret those tokens as

---

4 See Newell's Presidential Address to the American Association for Artificial Intelligence, reprinted in AI Magazine, 1981 under the title "The Knowledge Level".

having symbolic content, as standing for numbers and numerical operations. Since numbers and numerals are non-identical, we cannot identify a number with the numeral which represents it. The two are categorically distinct. But, for all practical purposes, we may think of the semantic characterization of the system as correlatable with a syntactic characterization. Hence, for Chomsky, theories about knowledge states, can be construed as theories about functional structure.

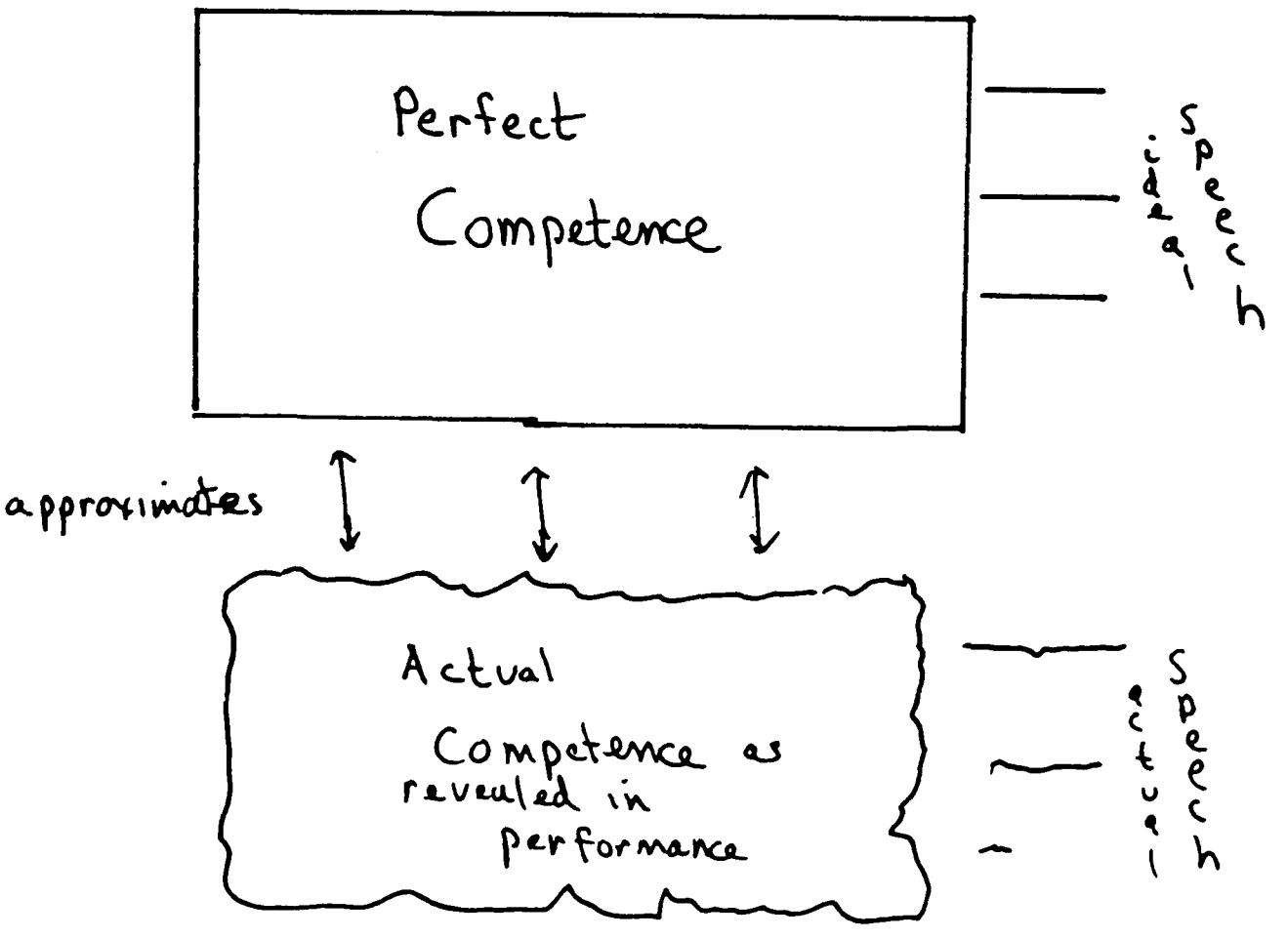
The position I defend departs from Chomsky's over the tightness of the connection holding between knowledge and function. Because of the directness of the image of knowledge as a proposition encoded in some inner language, Chomsky mistakenly, I believe, correlates knowledge states with functionally active sentences in the brain's "language of thought". Thus he seems to maintain, for instance, that the knowledge of the transformational rule involved in converting actives to passives, is correlated with some rule 'encoded' in an inner language, which contains symbols referring to phrase markers and phrase shiftings.

My own view is that theories of knowledge and theories of function are connected in less straightforward ways. It is idle to hope for neat correlations between knowledge states and functional states, even if we allow that those functional states are sentences in some inner language of the brain, for although knowledge constrains the sorts of functional processes that occur in cognitive systems, this constraining relation is far more complex than correlation. Indeed, it seems to me that the deepest problem facing cognitive scientists is to state the

nature of this constraint.

Confusion over the relation of theories of knowledge to theories of process tends to show up in most discussions of competence/performance distinction. In Aspects of a Theory of Syntax, Chomsky called the state of a person who knows his language his competence, contrasting this with his verbal behaviour and dispositions to verbal behaviour, or performance.

The common interpretation of the distinction, is that both competence and performance refer to abilities: competence, to an idealized ability which can be manifested in behaviour if the agent had unlimited computational power, memory and time, and which can be elegantly codified in terms of a system of recursive rules; performance, to the actual behavioural repertoire of the agent. At the very least, therefore, a theory of competence is taken to provide an elegant approximation of performance abilities. It represents the platonic capacity to speak which is imperfectly instantiated in material minds.



By interpreting competence as a description of the behavioural repertoire of an ideal agent, it is possible to see theories of knowledge as formal specifications of linguistic competence. In formal studies of specification[5] a system may be specified by means of predicates that do not directly refer to observable features of input or output, providing there is an effective procedure for translating sentences using those predicates into statements of observation. A specification, therefore, can abstract from details of particular interpretations. A formal specification may deal with highly general features of input-output relations that are invisible to theoreticians who confine their gaze to details of behaviour. Thus, if a system receives sequences of numbers as input and produces sequences of numbers as outputs, it may nonetheless be described as performing operations on sets, if it is easier to reason about sets than about sequences. In this way, theoretical discoveries drawn from set theory may then be exploited when it comes time to reason about the design of systems that might satisfy the specifications.

There are clear advantages to interpreting theories of competence as specifications of ideal behaviour, and theories of performance as specifications of actual behaviour. But such interpretations, though meaningful empirically, distort the distinction between states of knowledge and behavioural dispositions, violating Chomsky's deeper rationale for distinguishing them. Chomsky repeatedly spoke of a competence model 'embedded' or 'incorporated'[6] in a performance model:

---

5 See, for instance, the work of Oxford's Programming Research Group, as discussed in C.A.R. Hoare, "Specifications, Programs, and Implementations", 1982; Bernard Sufrin, "Formal System Specification" 1982, and his "Formal Specification of a Display Editor", 1981.

6 Op.cit. 1980, pp 200-201, 226, op.cit. 1972 p 117., op.cit.1965 p 10.

implying that knowledge of grammar forms the basis of behavioural dispositions in conjunction with other mental processes and structures.

There are two interpretations compatible with this view. The first, shown in figure 1.1, represents performance as the result of a cognitive processor reading a program powerful enough to produce ideal behaviour, but prevented from performing in idealized fashion by its own limitations. Fig 1.2, represents performance as the result of a cognitive processor reading a program that contains a theory of idealized behaviour, but prevented from causing that ideal behaviour because of the dampening effect of mediating processes. The two are

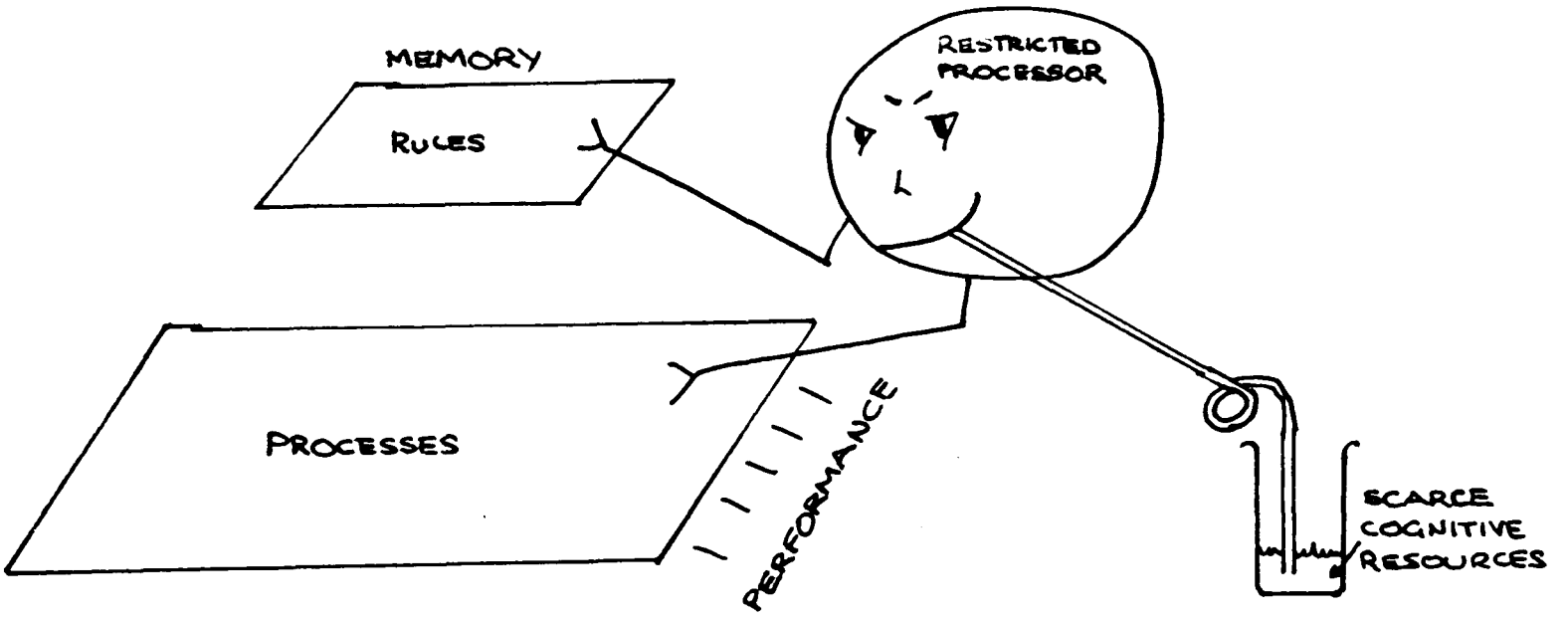


Figure 1.1

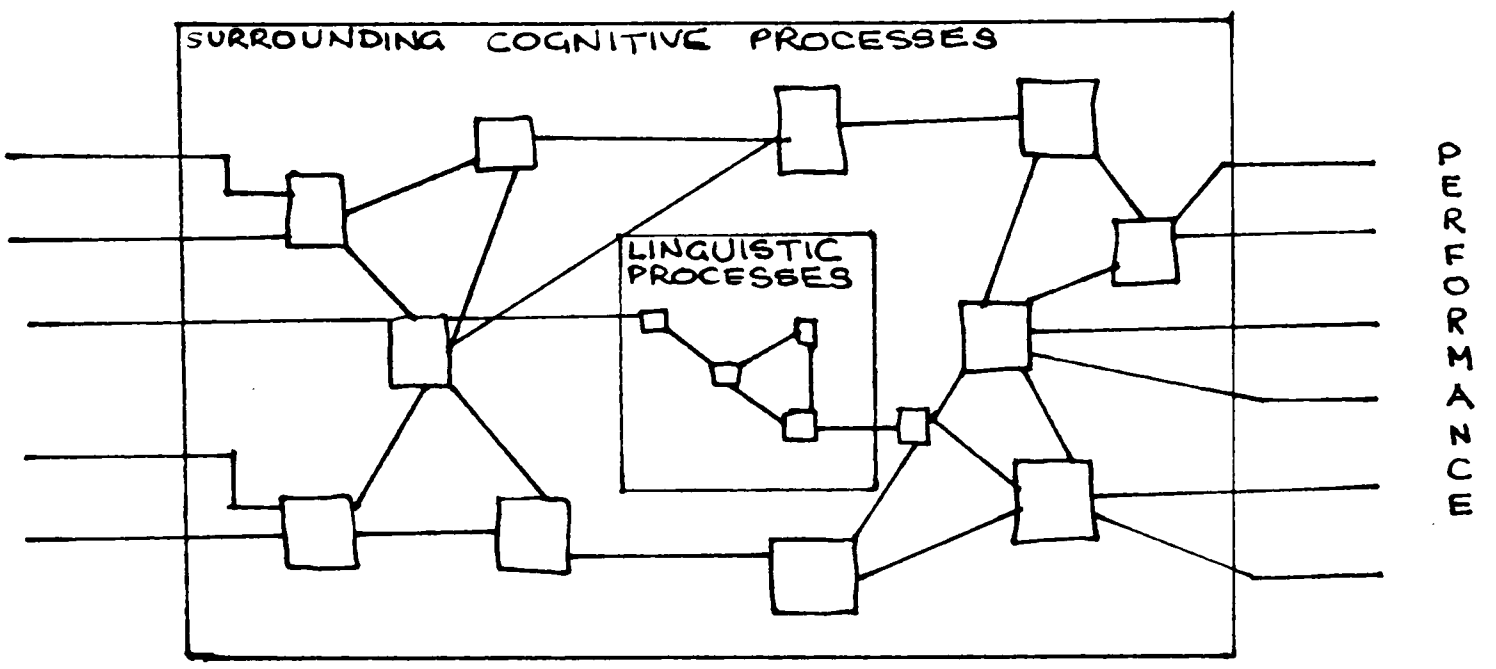


Figure 1.2

different, but share the key idea that states of knowledge associated with grammatical competence are real states of the system, logically independent from whatever behaviour may count as their manifestation. In each case, the agent has access to the rules of his grammar, but is unable to exploit that knowledge to the fullest possible degree.

If one rejects the correlation between knowledge and rules encoded in the language of thought, both interpretations of the competence--performance relationship lose their intuitive appeal. We are thrown back to the basic problem of stating how knowledge constrains process. As realists we feel certain that knowledge must somehow constrain the pattern of processes occurring in the system. As process theorists we believe that the pattern of processes determines performance. But what is the precise relationship between knowledge and process? Is it reasonable to hope that theories of knowledge can serve as guidelines in the search for process mechanisms? My objective is to show that it is.

Some Objections

There are many critics who disagree with the very terms of this problem. It has been argued, for instance, that propositional attitudes, and knowledge states in particular, are not real intrinsic states of a system. They are attributed, it is said, on the basis of a method of interpretation that is of questionable objectivity. Different observers can assign different interpretations of attitudes to a system, and there is no standard, even in principle, from which to adjudicate disputes.[7] Hence, propositional attitudes are spurious states; they may serve some purpose in the social and cognitive sciences when these sciences are in their early stages, but they designate nothing real in systems and so cannot possibly be a legitimate object of inquiry, and cannot possibly set constraints on system design or on systemic processes. Cognitive science, accordingly, should confine itself to the study of behavioural specifications, to the study of the properties of different computational processes, and to the way these two match up. For real cognitive science is a branch of control theory: neural control theory.

An opposite objection comes from those who grant the objectivity of the attributional approach, but think that it describes a system from a perspective that can shed no light on mechanical design. In

---

7 Foremost in this group is Quine. His celebrated Indeterminacy thesis is an undisguised attack on the legitimacy of propositions and hence propositional attitudes. Later developments of his position can be found in "Propositional Objects" appearing in Ontological Relativity and Other Essays. (New York: Columbia University Press. 1969) and in his replies to Davidson and Kaplan in Words and Objections: Essays on The Work of M.V.Quine, ed. by Donald Davidson and Jaakko Hintikka. (Dordrecht: D.Reidel, 1969).

this case, it is admitted that systems can be studied as receptacles of knowledge and desire; we can study cognitively endowed agents and legitimately ask what they know that allows them to perform a certain task. But cognitive science stops there. Describing an ability as a body of knowledge tells us nothing about the mechanical design of agents, or about the processes internal to them. The two frameworks, intentional and mechanical (structural) are different. They work with different conceptions of order. In the mechanical world the ordering principles are numerical. Laws are framed in terms of numerical functions, or numerical concepts such as 'X varies directly with...' . In the world of beliefs and desires, however, the ordering principles are semantic, laws if statable at all -- and philosphers, such as Davidson,[8] claim there are no laws at the intentional level -- describe connections between concepts and propositions. They exploit relations in the domain of propositions rather than relations in the domain of numbers. Hence orderings found in one framework cannot set constraints on orderings found in the other. The two are incommensurable. There is no informative relation to be found between knowledge and process.

The third objection comes from those who believe there is no principled distinction between cognitive and non-cognitive capacities. Since it is plausible to assume that intentional descriptions teach us about the structure of cognitive capacities only, we must have some

---

8 Davidson's views on the 'anomalous' nature of the mental are developed in "Mental Events", "Psychology and Philosophy" and "The Material Mind", all reprinted in his Essays on Actions and Events. (Oxford: Oxford University Press, 1980).

principled means of deciding what is a cognitive capacity and what is not. According to certain critics[9] no such principles exist. We are caught on a slippery slope between cognitive and non-cognitive and cannot draw the line needed to find constraints on processes.

The force of this objection may not at first be apparent. The thesis I shall be defending, however, is that knowledge is a real and irreducible state of certain organisms. Only if a system has an inner control structure of sufficient complexity to organise its information gathering, processing and storing so as to be able to 'think' about individuals in its environment, that is, if it is able to use concepts of objects and properties, will we be justified in asserting that it stands in a genuinely epistemic relation to its environment. As the cognitive system is a complex of faculties and capacities, some more central to the conceptualising apparatus than others, it is not always easy to delineate where genuinely cognitive operations begin and merely biological operations stop. Yet if we cannot draw this line between cognitive and non-cognitive operations virtually any biological process can be described as the result of a body of knowledge. Hence the response mechanism of an amoeba, or the orienting reflex of a mollusc, no less than the interpretation of speech by humans, would be describable as structured by states of knowledge. Obviously such a move vitiates the plausibility of the entire knowledge level research program. For if an ability, whatever its sophistication or its function can be called cognitive, then knowledge can be attributed to

---

9 Pat Churchland in "Neuroscience and Psychology: should the labor be divided?". The Behavioral and Brain Sciences. (1980),3, p.133.

any system, regardless of the complexity of its interactions with its environment. How then could knowledge set constraints on lower level processes? As the complexity of an ability falls, the diversity of systems able to display that ability explodes exponentially, decreasing the chances of discovering any structure to information processing which all or most systems must have.

I will have much to say about these objections in subsequent chapters. All can be met. The most radical, however, is Davidson's: theories of knowledge are incommensurable with theories of mechanistic processes. Accordingly, in the first chapter, I present my reasons for thinking that knowledge might constrain process. Even if one were to believe that there are, in fact, two orders in nature, he need not assume there are no relations between those orders. But as I argue in Chapters Four and Five, the extent to which we can expect to gather hints about the design of mechanistic processes by investigating cognition at the knowledge level depends on how complete our theory at the knowledge level is. In the domains of greatest interest to humanists -- that is, thinking, problem solving, rational choice -- the conditions necessary for building informative knowledge level theories are impossible to satisfy. Nonetheless, failure in one domain of inquiry is no guarantee of failure throughout. In Chapters Two and Three, therefore, I discuss the basic methodological assumptions of the knowledge level and the reasonableness of a moderate realism about knowledge states.

Chapter One

INTERLEVEL RELATIONS

Any complex system, whatever its composition, can be studied from different analytical frameworks, different perspectives. Philosophers typically distinguish structural, functional and intentional frameworks -- though there are certainly other ways in which analytical perspectives can vary. For instance, the level of abstraction or generality of an approach is another dimension along which analysis can differ, as is the spatio-temporal focus of the analysis.

In distinguishing frameworks one invites questions about their relation. It is noteworthy that the answers in philosophy so far have been polarised toward the extremes of reductionism or pure incommensurability. The objection raised by Davidson and others to the project of searching for constraints between knowledge and process is a case in point; it relies on there being virtually no connections to be found between intentional and structural or functional analyses, save the minimal conditions that intentional descriptions must be supervenient on structural or functional descriptions.

Since philosophers of science attempt to define interlevel relations in terms of the logical relations available in the predicate calculus -- logical sufficiency, logical necessity, biconditionality, and identity--it is hardly surprising that relations weaker than reduction yet stronger than supervenience have received little investigation. In this chapter I shall argue that the relation which

knowledge states bear to the processes which realize them falls in that logical netherworld. I begin the discussion with a brief historical account of the shifts in opinion concerning the relation of knowledge and belief to material processes. Then, to motivate a search for a different interlevel relation, I consider how function constrains structure in the theory of design. If the relation between function and structure is not as loose as modern tradition claims, perhaps the relation of knowledge to process is less loose as well. Nonetheless it is not as strong as correlation. I conclude this chapter with a criticism of the modern correlationist stance.

Some Historical Background

Before 1960,[1] the prevailing opinion among philosophers was that mental states, that is beliefs, desires, thoughts, and images, were either nomologically correlated with structurally defined states of the nervous system, or identical with them. There was not much empirical evidence for the claim, but at the time the hypothesis was not actually inconsistent with neurophysiological findings and it had the advantage of salvaging the unity of science. This last goal was the active motivation. It was readily allowed that science might appear, at any moment, to be a collection of unconnected theories and models. But it was thought virtually inconceivable that nature might be truly

---

1 I choose 1960 as the turning point to coincide with Putnam's first article on functionalism, originally published in Sidney Hook (Ed.) Dimensions of Mind (New York: University Press, 1960). The standard identity or correlation theories to which Putnam was proposing an alternative were advanced in the late 50's by U.T. Place in "Is consciousness a brain-process?", and J.J.C. Smart in "Sensation and brain processes"; both reprinted in C.V. Borst (Ed.), The Mind Brain Identity Theory, London: Macmillan Press. (1970).

disjoint: that the true theories of any field could not, in principle, be co-ordinated with the true theories of others. The view was typified (and still is) by Herbert Simon:

> We see nature as an immmensely complex hierarchical system, understandable only through being represented alternately at many levels of detail and understood by constructing bodies of theories at each of these levels of detail, in combination with reduction theories that show how the unanalysed elementary structures at each level can themselves be explained in terms of the constructs available at the next level below.[2]

Nature was seen as a vast Chinese box of systems, each system of properties constructed out of ones from the level below. Intentional Psychology, insofar as it had a legitimate place at all in science, sat near the highest level.

By the 1960's philosophers began arguing that correlations between intentional and structural states were improbable. Philosophical analysis of the nature of the interpretative process -- the method of ascribing beliefs and desires -- showed that what establishes the truth (or prima facie truth) of an ascription is behavioural capacity. A priori then, there is no reason why creatures without brains could not have beliefs and desires.[3] Thus if a non-human system shared all behavioural dispositions with a person it necessarily had as much claim to having mental states as the person himself. It seemed a simple step

---

2 Models of Thought (New Haven: Yale Univ. Press) 1979, p.63. My emphasis.

3 Again see Putnam: "Minds and Machines", "Brains and Behaviour","Other Minds", "Robots: Machines or artificially created life?", all reprinted in Volume 2 of his Philosophical Papers: Mind, Language and Reality, (Cambridge: Cambridge University Press, 1975). Also, see Jerry Fodor, Psychological Explanation, (New York: Random House, 1968).

to arguing that intentional states cannot be nomologically correlated with structural states of the brain. For as it is possible, in principle, to construct from non-neural parts control mechanisms that do the job of the brain, intentional states can be attributed to systems that lack neural tissue.

In a world where all control mechanisms have a physico-chemical basis, arguments against psycho-neural correlations are not strong enough to disprove psycho-physical correlationism. Although it is contingent that brains are the behaviour control mechanisms of all known mind-endowed creatures, it is not contingent, in a material universe, that all such control mechanisms are physical. Might there not, therefore, be correlations to be found between being in an intentional state, say, instantiating a certain field of beliefs, and being in a (highly abstract) chemical or physical state?

It was here that die-hard anti-reductionists played their trump card. The reason intentional states can never be correlated with structurally defined properties of systems is that they are individuated on the basis of radically different principles. Although we may state the content of intentional states in sentences, the content itself is not structured the way the sentence is. Two sentences of different grammatical structure may express the same proposition, as in 'John is a bachelor'and 'John is an unmarried man'. There is no structural property which synonomous sentences must share. What makes two sentences synonymous is that they mean the same. Structurally they may be arbitrarily different. Hence the principle by which we individuate mental content cannot be the same as that used to individuate spatio-temporally structured entities. How then could

thoughts and beliefs have structural counterparts in the physical world?

The pendulum had apparently come full swing. Abandoning a tight nomological correlation between mental and physical state, philosophers still convinced of the material basis of mind, invoked the concept of supervenience[4] to explain the connection. The Chinese box model of nature collapsed to be replaced by a less strict hierarchy of loosely coupled levels.

Supervenience itself is an extremely weak relation. A set of predicates P is said to supervene on a set of predicates Q if for any two systems x and y:

(1) if x differs from y with respect to some P predicate, then x necessarily differs from y with respect to some Q predicate. That is, a system cannot live through a change in its high level properties without simultaneously living through a change in some of its low level properties; and

(2) if x is equivalent to y with respect to all Q predicates, then x is necessarily equivalent to y with respect to all P predicates. That is, if two systems are identical at lower levels of description they are necessarily equivalent at higher levels.

Thus supervenience is compatible with emergentism: the doctrine that irreducible properties 'emerge' as we ascend the ladder of complexity.

---

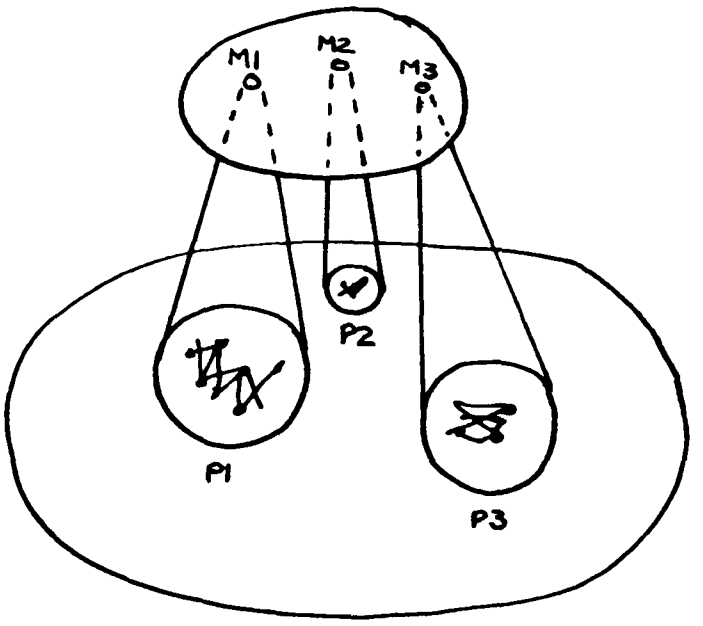
4 Although originally a term used in aesthetics and ethics by Moore and others to refer to the material basis of aesthetic and moral properties, it was re-instated as a catch-word in philosophy by Davidson in "Mental Events", reprinted in his Actions and Events, (Oxford: Oxford University Press, 1980).

In fact, we may think of it as emergentism in the formal mode.

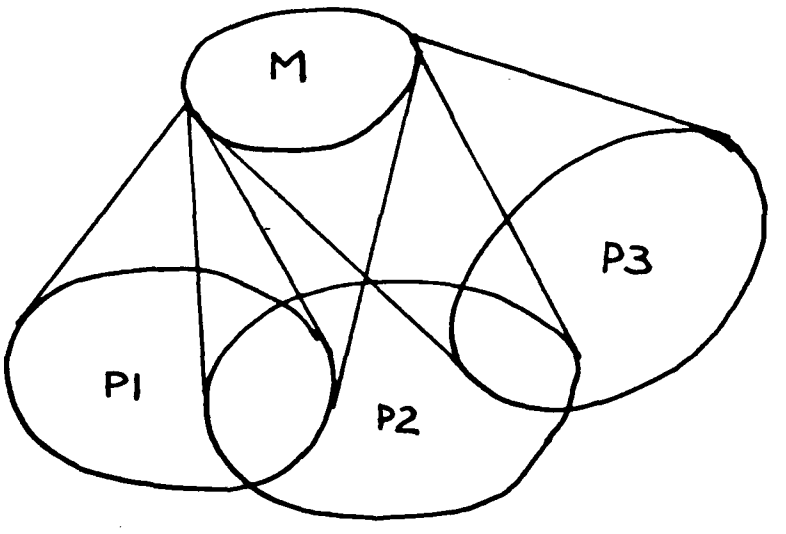
A good example is found in magazines with colour reprints. A picture of a person can be described at a high level of description in terms of human characteristics, a nose, a hand, clothes and so forth. It can also be described as a structured mass of coloured dots. The dots can be changed, within constraints, without changing the characteristics describable in personal terms. But it is impossible to change a personal characteristic without changing the dots. Thus a research scientist interested in differences between Japanese and American magazines could study differences in their publishing technologies, focusing on their techniques of colour reproduction, dot distribution etc., or he could study differences in facial characteristics, cartoon style, layout, setting and so on. The two studies could go on simultaneously and independently. Indeed, it would be surprising if there was much overlap.

At the level of dot technology it is improbable that we could specify which dots must be altered in order to, say, show a cartoon sequence of a samurai getting angry. The key relations in emotional representation concern relations between eyebrows, mouth and other facial features. Needless to say, all cartoon displays of anger presuppose sequences of dot changes. But there may be no single structural correlation. The study of dots and the study of human characteristics could be carried on autonomously. Neither one would encroach on the other. See Figure 1.1

Reduction vs Supervenience



Nomological correlation between mental states or processes and physical states or processes.



Supervenience of mental on the physical: any change in mental processes presupposes a change in physical processes, but not vice versa. Radically different physical systems may realize the same mental system. Indeed one can imagine a continuous changing physical system yielding a stable mental one.

Figure 1.1

Supervenience is Too Weak

From the armchair, supervenience seems exactly the relation needed to retain materialism while giving the Special Sciences -- cognitive psychology, linguistics and the human sciences -- the methodological autonomy they want to construct theories at the intentional level

unhampered by restrictions imposed from physics, chemistry and neuro-physiology. Yet the freedom cuts both ways. If Intentional Psychology can be carried on 'autonomously', it also cannot give any leads on structural order below. The search for mechanisms of behaviour would be undirected by studies at the intentional level.

Doubts about the necessity of describing the relation between mind and mechanism as one of supervenience ought to be aroused by the fact that engineers, architects, and computer scientists routinely 'infer' structure from function. The received view in philosophy is that functional states no less than intentional states supervene on structural states. In this respect, the relation of function to structure is analogous to the relation of mind to mechanism. Allegedly, there is no limit to the number of structures that may realise or subserve a given functional state. Yet, if that were true, how can inferences be made from function to structure as design scientists regularly do? There must be a path linking functional description to structural description. But if there is a path from function to structure may there not be one linking intentional description with functional description? If philosophers are wrong about the relation between function and structure might they not also be wrong about the relation between intention and function?

One of the awkward problems frustrating efforts to prove that the relation between function and structure (and correspondingly between knowledge and mechanism) is stronger than bare supervenience is that it is hard to define the precise logical nature of the relation which functional descriptions bear to structural descriptions. In the design

sciences we find maxims guiding design.[5] Evidently there are restrictions in the designing procedure. But a maxim is not a substitute for a definition: it relies on their being certain stronger-than-supervenience relations but it does not state them. To look for hints about the relationship between function and structure we might consider turning to studies on the logic of design. But again nothing is said about the precise relation of function to structure. Design constraints, it seems, are largely technology dependent. If we are told to design a car from existing technologies using standard parts and tools, then only certain designs need be considered. The class of possible designs based on those materials is severely constrained. But the restrictions come as much from the technology used as from the functional specification. Hence we cannot be sure that the same structural relations are required in other technologies.

From design theory alone we cannot prove that the relation between structure and function is stronger than supervenience. Nor is it evident that there are other sources which will provide the materials for such a proof. Yet although the relation is not provably stronger, the very fact that engineers can find technology-relative constraints constitutes an inductive argument for maintaining that function is more closely connected to structure than ardent anti-reductionists maintain. It would be foolhardy to suppose the function structure relation could resemble anything as strong as type-type correlation. Even the weaker relation of partial reduction: i.e. that all

---

5 See, for example, Blanchard, Benjamin and Fabrycky: Systems Engineering Analysis and Systems Design. (New York: Prentice Hall, 1981).

functionally equivalent systems share a certain structural attribute, which is necessary but not sufficient for functionality, is clearly too strong. Partial reduction would obtain if we could prove that no device designed to transport people could have less than ten moving parts, or that no device for solving the Travelling Salesman problem could fail to have parts that are isomorphic with the configuration of cities the salesman is to visit. But even if such proofs were available, they would be open to doubt on the grounds that they appear valid because of our impoverished imagination. It is hard to put limits on the possible. Even partial reduction, therefore, is likely to be stronger than we can expect.

From the difficulty to fit the relation of function to structure into simple logical notation, however, it still does not follow that the relation is not stronger than supervenience. Function constrains structure in some way. Our problem is that if we confine ourselves to the standard formalisms of logic, we cannot state how. By way of providing inductive support for the claim that this stronger relation exists let us examine an instance where we can calculate some of the design constraints which a functional specification imposes.

Design Constraints: an Example

Imagine we are asked to design a simple mechanism to be situated on the earth's surface which can accelerate objects in accordance with some function $v_2 = f(v_1)$. We are told that the mechanism is to be located inside a box whose dimensions are X by Y by Z, and also that the object entering the box will come in through an input hole at the top corner of the box and leave by the opposite corner on the bottom.

Can we make fairly general claims about simple designs that could not possibly work whatever materials or parts they are implemented in? Better yet, can we make any general claims about properties of designs that would work?

Without an adequate metric of simplicity it is hard to decide what constitutes a simple design. However as we know that the system must have some means of directing the ball from input hole to output hole -- a requirement we infer by restating[6] the functional specification in more structure revealing form — we may start our inquiry by considering the simple mechanism inside the box to be a slide. Already several consequences follow. Since we know that gravity is already acting on the ball, and we know the net vertical drop between the holes, we can calculate the contribution to velocity which gravity alone causes. Call this g. Because we know as well that this function g, is the maximum unaided acceleration an object will acquire under the stated conditions, we know that if the acceleration function f in $v_2 = f(v_1)$ is greater than or equal to g, we can infer that there must be something inside the box contributing to the acceleration of the ball. A slide could not do it alone. Of course this is not a great discovery. It doesn't take much ingenuity to infer that there must be a mechanism inside the box that accelerates the ball. Nonetheless the discovery sets a negative constraint on design. We have discovered that the mechanism is not just a slide. On the other hand if f is less than g, and we have an index of the friction different materials

---

6 For a tentative methodology for redescribing functions see Freeman, P. and A. Newell, A Model of Functional Reasoning in Design, CMU-CS-107, 1971.

present, we can predict the number/length of the intermediate up and down segments in a slide made from each substance. That is, we can produce a family of constraints.

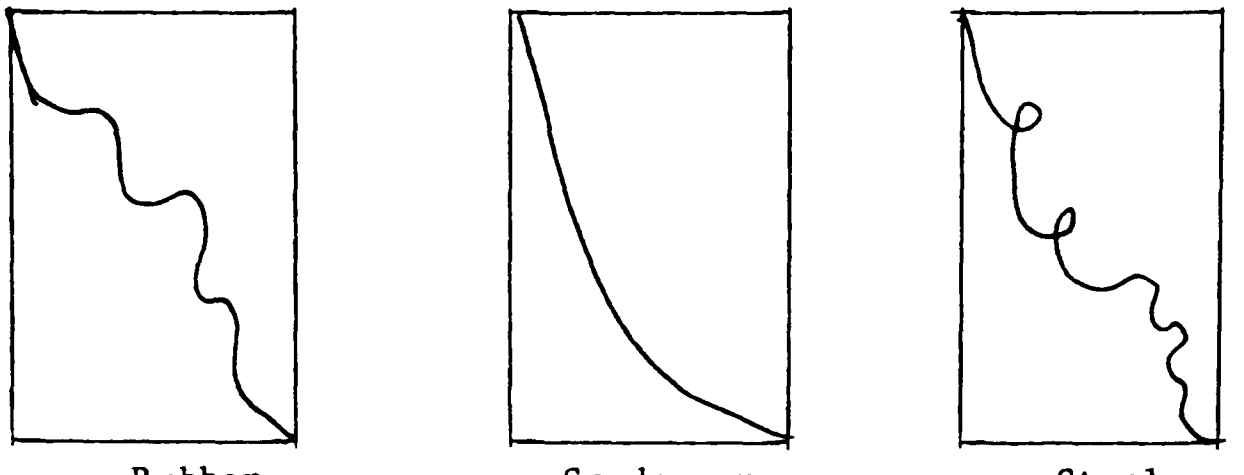


Figure 1.2

What have we achieved? We have found that given a perspicuous functional specification we can begin speculating, in a principled way, on the sorts of structures that might implement it. To be sure, the process requires research simultaneously on properties of the implementing medium and on properties of the system qua functional device. Our equation is that

Function + Technology = Constraints on structure

so knowledge of technology, no less than knowledge of function, is required to infer constraints. The importance of cooperation between technical engineers here must be fully appreciated. We concluded that if a substance has a certain coefficient of resistance then we can, on the basis of functional specifications, infer that <u>if</u> the design is a slide -- an assumption that is made on grounds of simplicity -- then there will be a calculable number of square inches of contact between the ball and the slide. The manner in which these inches of contact are achieved -- whether through many up and down steps, a few large

ones, or actual 360° loops -- is a further hypothesis. Function and technology jointly determine an equivalence class of designs.

Of course, we have said nothing about a property that all mechanisms meeting the specification will share. Given the possibility of insanely complicated Heath Robinson contraptions, it is probable that there is no structural property common to all members in the class of functionally equivalent devices. And yet for simple, efficient mechanisms genuine constraints on design have been inferred. Function and efficient structure are connected in some orderly fashion. It would be a mistake to think that the one is only supervenient on the other.

It was this observation that led me to suppose that intentional specifications may be structure-revealing too. There is no chance, of course, of finding correlations between bodies of knowledge and structural configuration in mind-like systems. Even if all mind-like systems operate with an internal 'language', an inner notational system, that 'represents' or encodes knowledge, there is still an indefinably large class of structurally distinct notational systems that might do the job. As with the function/structure relation even partial reduction is too strong. Technology specific -- that is notation specific -- constraints are all we can hope for. But if these constraints **can** be discovered we may then discover design maxims that will allow us to move from knowledge specifications to possible process specifications. Between task analysis and computational analysis there may be a third form of analysis, knowledge analysis. It will have its own methods, its own standards and its own biases. But it will issue in descriptions of states which are connected with lower levels in an orderly way. The question I shall be asking in this thesis:

Is it reasonable to hope for a science built on a research program that involves inferring mechanisms from knowledge specifications?

The Formality Condition

A number of attempts have already been made to state the sort of downward constraints that might flow from intentional specifications. The most famous of these, the formality condition, or more precisely the syntactic mirroring hypothesis, was introduced by Jerry Fodor in The Language of Thought,[7] and Zenon Pylyshyn in 'Computation and Cognition'[8].

The formality condition states that all symbol manipulations apply to symbols in virtue of their form, not their content. A system whose internal state changes are best described as a trajectory of changes in semantic states, never literally has access to the semantic interpretation of the symbols it operates with. If the system is mechanistic, it is a closed system in which state changes are determined by local interactions. Semantic interpretations depend on relations to objects outside the boundaries of the system, and therefore have no direct effect on internal change. One may assume that the system has been adapted to its environment so that it behaves as if it genuinely interprets environmental states. But its state changes conform to this 'interpretation' because certain formal symbol structures are operated on by formal rules in the right way.

The syntactic mirroring hypothesis strengthens the formality condition. Briefly put, the condition states that for each proposition

---

7 The Language of Thought, (Cambridge: Harvard University Press, 1975).
8 Behavioural and Brain Sciences, 3, pp.11-169.

a person believes (where 'belief' is treated as shorthand for any intentional state), there must be a formal counterpart in the brain, a network of forms, whose structure is isomorphic to the semantic structure of the proposition believed.

Beliefs have content: they represent the world or features of it as being such and so, and the causal role they play in a system is related to this content. Put differently, beliefs have a causal impact on a system commensurate to their sense. Unless there is action at a distance, however, the causal power of a belief cannot literally come from what it represents: the reference or content of a belief usually lies outside the system, in a proposition, fact or state-of-affairs. Somehow the properties of the state-of-affairs, must be converted into causally efficacious features inside the system. The only way this can be done, Fodor suggests, is if the structure inside the system which represents a given state of affairs serves as a symbol whose formal features mirror the features of the state-of-affairs. In that case, each explanation-relevant feature of a content-specifying state-of-affairs could be correlated with an internal causal feature. Having a belief would consist in being in a 'believing' relation to an internal form, an internal representation. Changes of belief, and so forth, could be explained by finding computational principles which apply formally to representations.

It is not hard to see what Fodor has in mind. If a system really has beliefs, and if its behaviour is explicable because it has them, then it may seem reasonable that there must be something inside the system, something underpinning the belief, which carries the same causal power as the belief. This inner something is what makes it true

that the system has the belief. How else could there be a genuine difference between a system which seems to have a belief and one which really has one? Whatever this inner something is, however, it must be able to enter into relation with other beliefs in the right sort of ways.

If there are systematic relations between beliefs, they must be mirrored by systematic relations between their internal correlates. That is, if the belief that p is deductively related to the belief that q, the internal correlate of the belief that p must be related to the internal correlate of the belief that q in such a way that operations corresponding to what we call deduction if performed on p and q can be performed on the correlate of belief p to derive belief q. The naive solution is if the internal correlates of the belief that p and the belief that q are embedded in a computational system and the formal structures and structure-transforming operations inside the head can be thought of as modelling propositions and proposition-transforming operations. The formality condition merely summarizes this point: for as modelling requires an isomorphism between the structures and relations modelled and the formalism doing the modelling, there must be a one-to-one correspondence between the elements and relations in propositions and the elements and relations in the internal formalism.

If true the formality condition imposes a fantastic constraint on the lower level design of systems. Every time we explain a system's activity by describing its reasoning, that is, by its trajectory of thought, we will know ipso facto that there must be an isomorphic trajectory of formal representations. Furthermore, were we to discover that certain cognitive abilities require the agent to know

certain facts or principles about its environment of action, we would also be able to infer that those facts or principles are actually represented in the agent. Intentional states could be nomologically correlated with formal states.

This is exactly the sort of claim I doubt we can make. Intentional states are not nomologically correlated with formal states. Even if we were to know an agent's inner notational system, we still could not infer what representations he uses to encode his knowledge and belief because there are simply too many different kinds of computational architectures that could use that notation to make it plausible to expect a neat correlation. Given the endless ways notations can interact with interpretors, correlations between intentional states and notational structures will be open-ended. This is not to deny that there is some informative relation to be found between intentional states and functional processes in a computational mechanism. But it will be far more complex than is allowed in the syntactic mirroring hypothesis. I will devote the rest of this chapter to defending these ideas.

The Space of Possible Mechanisms

There is a danger in stating on a priori grounds how things are. What looks necessary from one perspective may seem contingent from another. In the design science the warning is real. Our present understanding of the space of possible mechanisms is so limited we should not hope to find universal constraints on intentional systems. The formality condition, however, promises universality: whatever the mechanism, whatever the substance, if it

truly instantiates intentional states then

1) it has an internal notational system, a language of thought;
2) it operates according to computational rules;
3) its syntactic trajectory mirrors its intentional trajectory; and
4) it explicitly encodes all its propositional attitudes.

Although there are grounds for doubting each of these claims I shall confine myself to reasons for thinking that (4) applies to no more than to a fraction of computational systems. The formality condition is just too strong.

If an intentional system has an explicit representation as part of each of its intentional states then

A) for every element in the facts and relations that are known, believed or desired, there is a corresponding element in the system's internal language;

B) the syntax of the inner language allows the formation of internal expressions that are at least as complex as the facts or relations they 'represent'; and

C) every element in the inner language which 'refers' to an element in the task environment 'refers' to only one element.

This last point follows because if some of the terms of the inner language are ambiguous the same formal structure would be forced to play as many causal roles as states of affairs it represents -- a possibility Fodor explicitly disallowed.

To disprove Fodor's claim that an intentional system, if computational, must have an explicit representation of each element of the state of affairs it has beliefs about, it will be sufficient to show that consequences A, B or C are not true of all computational systems that are input-output equivalent with ones to which we correctly ascribe beliefs and desires. In that case, we will have

found computational systems which do have beliefs and desires but which do not have internal representations of each belief.

Example 1: Procedural vs. Declarative Representations

The first example I have in mind derives from an obvious fact about commands: you do not have to say 'Don't do X!' if you can command a person to 'Do Y!' (where Y ≠ X). Nor must you say 'Do X' if you can command her to 'Act! But don't do A, or B, ...'(leaving only X to be done). This may seem a trivial point, but it illuminates a distinction between explicitly commanding (programming) a system to produce some effect and implicitly programming it.

It will be recalled that the basis for ascribing intentional states to a system is performance. A system operating in a certain environment is assumed to have certain goals. We study its behaviour to see if it undertakes actions that are rational given those goals. The more rationality we find in the action the more perfect we assume the agent's knowledge of its environment. In one respect this is perfectly justified. A rat that steers clear of dead ends in a maze knows that dead ends lead nowhere desirable. If it knows where food is usually located and where paths lead, or perhaps do not lead, its directness is understandable; it wants the food as soon as possible, following the shortest path is the way to get there, it knows which is the shortest path, so off it goes.

Now assuming our rat is artful, and it displays a skill and persistance in food finding that requires attributing to it cognitive processing, are we obliged to assume that it actually represents paths that are dead ends? Most psychologists to-day would agree that

navigation (certainly in humans) is 'based on processes of inference within the structure of a cognitive map.'[9] Suppose we grant the rat this talent. Are we obliged to say that the rat also has a representation of the fact that the shortest path is the least time consuming? The rats' goal was specified with respect to time: it wants to reach the food as soon as possible. The path it chooses is the shortest in length. We seem forced by our attributional method to ascribe the rat knowledge that shorter paths are briefer. Yet does it represent this knowledge? Might it not simply follow a program that enjoins it to avoid dead ends? In many environments avoidance of the bad leads one to the good. The beliefs and rules encoded may be about bad states, leaving implicit beliefs and rules about the good. At the intentional level, we have no easy means of distinguishing those beliefs; no way of knowing whether the system knows the good, intentionally pursues it and accidentally avoids the bad, or whether it knows the bad, intentionally shuns it and accidentally pursues the good.

The point I am making is that an agent may adapt to an environment by exploiting certain redundancies, regularities or structural relations but not actually represent these environmental features internally. Of course, the agent could not then reason about those features: having no representations or conception of them, he could not ask questions about them. But lack of explicit knowledge is not evidence that he does not implicitly know them. Implicit knowledge may

---

9 Oatley, K.G. (1977) "Inference navigation and cognitive maps". in Johnson-Laird and Wason (Eds.) Thinking: Readings in Cognitive Science.(Cambridge: Cambridge University Press, 1977),p.537.

be a real state of the system for all its lack of an explicit representational underpinning.

Another example may bring this closer to home. In the AI journals in the early 70's there was a considerable stir over the comparative benefits of declarative versus procedural representations. A procedure is a small program, it is a recipe for action, a set of commands. Knowledge is 'embedded' in programs using procedures but often not explicitly stated. A declarative, by contrast, is a straightforward assertion of the fact or relation known. It is utterly explicit. Procedures have their place.

> If we want a robot to manipulate a simple world (such as a table top of toy blocks), we do it most naturally by describing its manipulations as programs. The knowledge about building stacks is in the form of a program to do it. Since we specify in detail just what part will be called when, we are free **to build in** assumptions about how different facts interrelate. For example we know that calling a program to lift a block will not cause any changes in the relative positions of other blocks (making the assumption that we will only call the lift program for unencumbered blocks). In a declarative formalism, this fact must be stated in the form of a frame axiom which states something equivalent to 'If you lift a block X, and block Y is on block Z before you start, and if X is not Y and X is not Z and X is unencumbered, then Y is on Z when you are done" This fact must be used each time we ask about Y and Z in order to check that the relation still holds. Note that this knowledge is taken care of "automatically" in the procedural representation because we have control over when particular knowledge will be used[10]

Winograd's point is that if we organize a system so that it performs actions in a certain order we can **build right into that order** implicit knowledge of the environment. The system doesn't have to explicitly represent the law that moving one set of blocks will

---

10 Winograd, "The Procedural Declarative Controversy", in Bobrow D. and Collins, A. (Eds.), Representation and Understanding: Studies in Cognitive Science. (New York: Academic, 1975), p.189.

leave undisturbed blocks unconnected to the set, or that two blocks cannot occupy the same space-time region. Its actions are elicited in sequences that never make it try to force one block into the space of another, or force it to engage in costly observational searches to determine if untouched blocks are still there. Knowledge of these principles is built into the mechanism transferring control, as some say, and not in the facts represented. Consequently key environmental facts and relations can be exploited without being represented.

Does the system believe that blocks are impenetrable? If it has such beliefs they are at best implicit. At the same time, if it did not hold those beliefs, there would be no good reason to assume that its actions would continue to reflect environmental constraints. Generalisation about its conduct would lack counterfactual force. There is some justification for ascribing belief. I will return to this problem later.

The parallel with commanding an agent to do B so that it will not do A ought to be clear. The agent acting as commander knows the relevant facts about the environment, just as the designer of Winograd's robot knows about blocks and space. He builds that knowledge into the system by carefully stating commands whose execution respects environmental relations. But look anywhere you like, data structures, program rules, operating systems, you will find nothing explicit in the machine to correspond to the knowledge 'blocks are solid objects', 'two blocks cannot simultaneously occupy the same space' and so on. Intentional states justifiably represented at the knowledge level need not be explicitly represented at the computation level.

Example 2: Ambiguity and Ellipsis

The example above was to show that there need not be an element in a system's internal language to correspond to every element in the facts and relations it knows. Condition A is false. I now want to show that Condition C is also false. It is not the case that "every element in the inner language which 'refers' to an element in the task environment 'refers' to only one element."

It is a consequence of the formality condition that "two thoughts can be different only if they can be identified with relations to formally distinct representations". Representations must be unambiguous; they cannot 'mean' one thing in one context, another thing in a different context. Otherwise different thoughts might contain the same formal representation. Fodor's motivation for this rather strong requirement is that "to put it mildly, it is hard to see how internal representations could differ in causal role unless they differed in form". Why? English sentences can be ambiguous and we machines succeed in treating each different sentence meaning appropriately.

The real rationale, I think, is that Fodor assumes that the mechanism 'interpreting' the rules, the processor, or, if this sounds overly anthropomorphic, the body of dispostions inherent in the functional architecture of the system, is too simple to have context-sensitive dispositions. Although this may be true for dispositions construed as simple stimulus-response connections, not all computational models need rely on such a simple notion of disposition. Everything turns on how smart the processor is. If the processor has its own memory, it will have among its dispositions a tendency to disambiguate representations by their context.

Let us represent a computational system as a tripartite system. This is not a conventional Turing machine representation but is provably equivalent.[11]

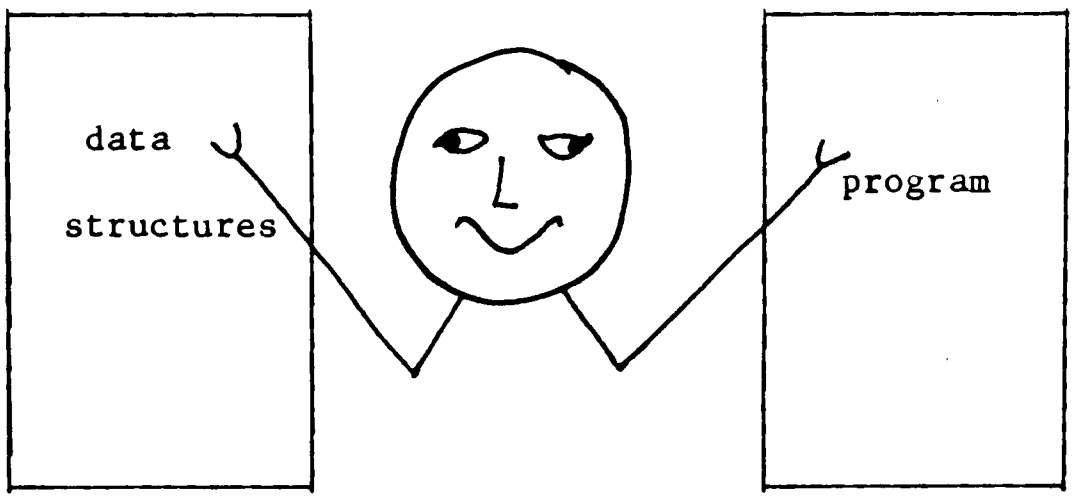


Figure 1.3

In Figure 1.3 a computational system is shown as having three logically distinct parts: a set of data structures or declaratives, written in language L, a program written in a logically distinct language L*,[12] and an interpreter, whose job is to interpret the instructions in L* as operations over the structures in L and carry them out. The representations in the data base and program are all static structures. They only produce computations when there is an active device to breath life into them, to create interaction. In

---

11 cf. Allen Newell, "Physical Symbol Systems", Cognitive Science, 4, esp. Sections 2 and 3, pp.142-54.

12 Computer scientists do not distinguish the language L, in which data structures are written, from the language L*, in which the program is written. Their justification is that what serves as a statement in a program can be introduced as data. Notionally the languages are the same, and it is often possible, as in LISP, to have statements of any type refer to statements of any other type. From a logical point of view, it is desirable to regiment language into types. Since programming rules are defined over data structures, prima facie they belong to higher order languages than those structures. One of the few members of the AI community to acknowledge this need is Brian Smith. In "The Computational Metaphor" (unpublished manuscript), he has argued forcefully for the importance of noting metalanguage/language distinctions in programming languages.

normal computers the processor is itself programmed (usually in a formally different language L**). For the behavioural dexterity it requires is often so great that the easiest way to achieve it is by programming it to respond 'appropriately' to the presence of rules and data. If we wish we can program our processor to be arbitrarily 'clever'.

Now given a system with such flexibility, it is easy to prove that the same formal structure — 'bear' or 'bank', for instance -- can be interpreted differently by the processor depending on its context. Terms that are formally similar in the language of thought pose no problem for a smart processor. It can interpret them differently according to their location in the system or to the context in which they are used. If the processor actually parses the sentence, for instance, it can distinguish 'bear' the verb, from 'bear' the noun. If it scans for semantic well-formedness, it can distinguish differences between the lexical meaning of 'bank' in 'We held the party in the bank' and 'We held the party on the bank'. Moreover a rule of the form $R_1 \rightarrow R_2$ can be interpreted by the processor as $R_1 \rightarrow R_2$ in one context and as $R_1 \rightarrow R_3$ in another. The processor can interpret the rules differently because it interprets $R_1 \rightarrow R_2$ as shortform for a context sensitive rule. Yet if ambiguity can be handled just by increasing the power of the processor, why make Condition C a universal claim about all intentional systems? The smarter the processor, the less problematic is ambiguity. Evidently Condition C is too strong. Therefore formal elements in the inner language can be assigned more than one semantic interpretation.

A similar argument can be produced for interpreting elliptical data and rules. By raising the intellect of the processor it can cope with implicit rules and data. A phrase or sentence is elliptical if part of its syntactic structure is implicit, to be added to the structure by its interpreter to make it syntactically well-formed and semantically meaningful. In the sentence

Do you take this woman to be... . I do.

We use factual knowledge to fill in the '...', and we use knowledge of linguistic context to interpret 'I do' as 'I do take this woman ... .'. Where there is ellipsis there is less structure in the representation than in the thing being represented. The rest of the structure is implicit. Hence if a computational system can work with elliptical data or rules it need not explicitly represent the structure of a state of affairs to display knowledge of that state.

I think it is fairly obvious that a smart processor can handle elliptical rules and elliptical assertions. But if so, there must be 'thoughts' such a system could have which were not completely represented in either the data base or rule base. To be sure, this does not prove that the syntax of the inner language is <u>incapable</u> of generating the structure of the state of affairs or action. But we can imagine concrete examples where a smart processor can compensate for an impoverished syntax and manage with an inadequate language of thought.

If these fanciful ideas about smart processors are true we have a patent violation of the claim that every intentional state is explicitly represented in the language of thought. In both the cases we are envisaging, the relevant lexical structures are strewn about the

system, some in the language L, some in L* and some in the language programming the interpreter. It cannot be necessary for intentionality that the syntax of the inner language, L or L* allow the formulation of internal expressions that are at least as complex as the facts or relations they 'represent'. There are many ways of compensating for an impoverished language of thought. Condition B is too strong: there is no saying, a priori, how complex a system's notational system must be.

Learning

I now want to take a different course. We have seen that none of Conditions A, B or C are necessary for intentionality. A system may have beliefs and desires but fail to explicitly encode each of those beliefs in a sentence in the language of thought. Nevertheless, it might be argued that Conditions A, B and C are plausible for any system that can learn. Thus although we can imagine computational architectures which simulate the behaviour of rational agents in static trials, despite relying on an impoverished language of thought, we cannot imagine them as systems that can learn. Hence systems able to learn always encode their beliefs explicitly in rules or representations.

Again I think this is demonstrably false. The possiblilty of much of the intelligence of a system residing in the mechanism which controls the use of rules and representations provides a reason for suspecting that **learning** may occur through changes in the interpreter, it need not be confined to local changes in data structures or procedures.

I mentioned earlier that a computational mechanism can be viewed from two perspectives. It can be seen as a formal system marching through structural state changes because of formal structures interacting with formal structures; or as a model whose inner structures correspond to structural states in some intended interpretation, say a task enviroment, and whose state transitions correspond with actions, operations or laws that can alter environmental states. Normally rules perfectly correspond to actions or operations, but as I suggested, this is not necessary: a rule need not have as much structure as the environment transforming action it is related to; the connection between rules in a program and state transition in a task environment may be extremely hard to state. Similarly in simple systems data structures perfectly correspond to outer structures. But again this is not necessary; the data structures inside a system need not have as much structure as state structures outside. The relation between data and environment may also be complex. We can, if we like, abstract from these complications and pretend that the system we are studying is a simple one, with a simple processor, and explicit representation of all environmental states and actions. In so doing we make it easy to interpret a computer loaded with a data base of 'facts' and 'relations' as an intelligent agent. We can see it as simulating, say, a medical diagnostition. For its input and output will relate to medical questions and answers and its state transitions will relate to the doctor's reasoning process. This picture satisfies the formality condition.

One attraction of this picture is that it makes it easy to frame questions about how a system could increase its body of knowledge. If

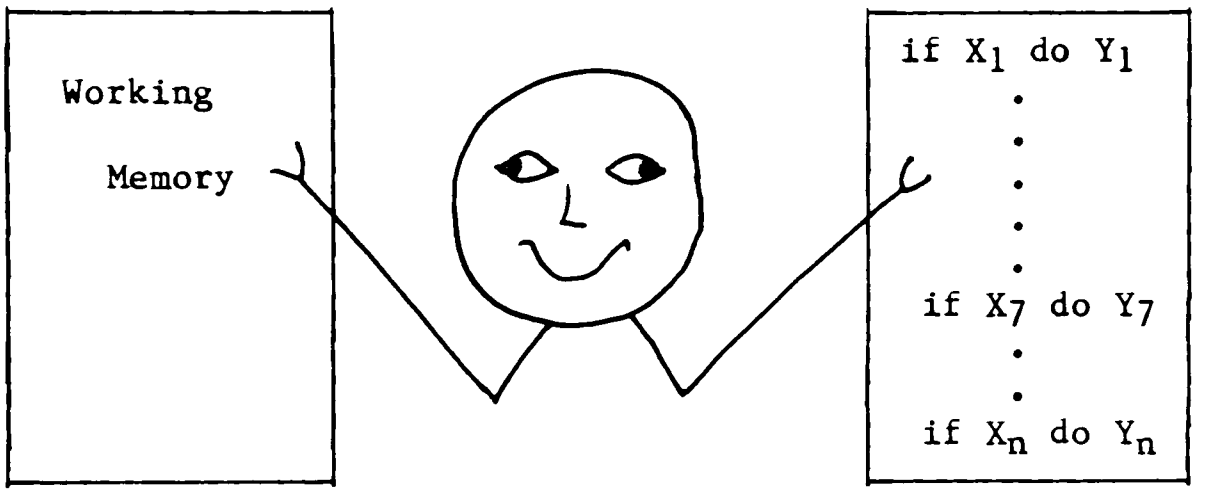
knowledge resides largely in the data base, and the procedures found in the system are correlates of the rules of deduction, the consequences of adding a new 'assertion' to the data base can be routinely calculated. Hence if the knowledge base is a set of independent 'facts', it can be altered in a predictable way simply by adding or subtracting assertions. The modularity of 'learning', in a trivial use of that term, can be explained. Similarly for procedures: modularity can be built up if all rules interact with data in a limited way (e.g. production systems). It is still hard to predict the consequences of a change of rules -- for minor changes can have far-reaching consequences -- but the model is clear.

Now, I do not want to suggest that Fodor lays undue stress on this simple model as a way of explaining learning. But it is clear from his writing that he considered most of learning to follow a hypothetico--deductive path, and that is exactly the line of thought most readily interpretable in 'here is the data, there is the program' language. Thus, insofar as learning seemed potentially explicable in this framework, it supported his formality condition. But by the same token, any reason to think that learning does not occur through explicit change in a 'rule' or 'fact' encoded in the system, is also a good reason to think the formality condition is false.

There are at least two ways in which a system might learn, yet which do not fit easily into the simple change-in-data/ change-in--program metaphor. The first plays on the idea of an interpreter itself being programmed; the second relies on models of parallel processing.

Production Systems

Among existing computer architectures production systems are noteworthy for making explicit the role of the interpreter. A production system consists of a set of procedures called production rules, a data base or working memory to which these rules are applied, and an interpretor which brings the productions into contact with memory and decides which production is to be executed at any moment[13]. In a production system there is no program per se. There are just production rules and a control regulating when they are activated



Production systems have been seen as a promising architecture of the human mind because they are modular, and we can imagine a creature learning simply by gaining or losing a few productions. What is seldom considered though is whether the control might change. In normal cases the function of the control is trivial. In the simplest execution schemes it checks the conditions of each production against working memory which holds the current state of the data base. It

---

13 For a lucid account of production systems and their role in Computer Science, see R.M. Young, "Production Systems for Modelling Human Cognition", in D.Michie (Ed.), Expert Systems in the Micro-Electronic Age. (Edinburgh: Edinburgh University Press, 1979); and Allen Newell, "Production systems: Models of control structures". In W.C. Chase (Ed.), Visual Information Processing, (New York: Academic, 1973).

identifies all the productions that might fire and then ranks them on the basis of some ordering principle. The orderings which have been most carefully explored are simple linear orderings of the productions (e.g. by recency of last use). If the control can find no production to activate it halts. If, after it has already ordered productions, it still finds more than one production capable of being fired, it assumes an error has been made.

Given the vast number of alternative orderings production systems might use it would be a surprise if there did not exist systems which could learn, in certain domains, by tampering with their control. Weightings, biases, even high order strategies, could be altered in the control, in an environmentally controlled way, that might, in certain cases, prove adaptive. For instance, suppose that in thinking about a certain topic I invariably become aroused emotionally and begin to change my way of thinking. The causal mechanism responsible for tying my emotions to the contents of my thoughts, for all its complexity, may nonetheless be acquired. It may be a learned process. Yet the result of this learning is that I live through a change in the form-manipulating dispositions of my reasoning system and acquire new sorts of beliefs because of it. In the grip of emotion I may 'grasp' inner symbols differently. I am inclined to draw different conclusions. I think differently. Yet these inner changes -- the fruits of my learning -- if well-defined at all, are changes in my dispositions that are to use production rules. I may retain the same production rules, syntactically defined, but change my way of interpreting them.

Now I am not suggesting that the chances are high that humans do learn in this way. But the possiblity should not be ruled out a

priori. La coeur a ses raisons que la raisonne ne sait pas. Certainly learning in computational systems, taken as a class, does not presuppose the formality condition. Hence the formality condition is not defensible on the grounds that only a system with a language of thought encoding all beliefs would be capable of systematically acquiring new beliefs.

Parallel processing

Parallel processing systems are a second possible computational system that might be capable of learning. There is reason to believe that they too violate the formality condition. A parallel processing system is a complex computational device made up of a community of computational machines. Although little is known about the methods of organising parallel systems it is at least known that regularities in interaction can emerge without being programmed. Five 'agents' each operating on its own task might reach a point in their computations where each sends the same signal to all others. This may recur. Cyclic patterns may emerge in community wide signalling. It is not impossible, that what we call belief is connected to the community wide states of parallel computational systems. Like a slot machine whose cogs all run independently, our brain may have separate computers each operating in relative autonomy but whose interactions determine our behaviour. The possibility cannot be ruled out that such systems might learn. Nature has found indefinitely many ways of achieving adaptiveness. Why not through a community of inner agents? Yet if parallel systems may acquire new global dispositions, learning may occur without sentences being encoded in the language of thought.

Learning would reside in the changed dispositions of the inner community to interact. There may be no local features of any one processor to identify with the cause of the new dispositions. So once again a system violating the formality condition might learn.

Conclusion

Unless all these arguments are unsound it follows that the syntactic mirroring hypothesis is too strong. There are computational alternatives to simple program/data architectures which resist it. This is not to say that homo sapiens does not have the right functional architecture to satisfy the formality requirement. But any constraints which intentional specifications impose on the design of systems must be discovered by empirical research. They are not *a priori* discoverable by philosophical inquiry alone. It is largely an empirical question as to what constraints flow downward from the intentional. These may vary from functional system to functional system.

Chapter Two

CENTRAL CONCEPTS

In this chapter I shall analyse two of the three major concepts of knowledge level research: task environment and task knowledge. It will be recalled that I have been arguing that cognitive research ought to proceed in two steps. In the first step cognitive systems ought to be interpreted as rational beings, endowed with knowledge of their environments, and motivated by goals and desires. Questions about the modularity of knowledge, sufficiency or necessity of knowledge, ease of knowledge acquisition and so forth are addressed at this stage. In the second step, inferences are to be made about the design of control systems that might instantiate those kinds of knowledge states, goals and rationality. Studies at the knowledge level serve as guidelines in the search for mechanism.

My case for structuring cognitive research in a top-down manner would be considerably strengthened if we had some paradigms of knowledge level research. In the penultimate section of this chapter I shall discuss David Marr's work, which I think gives us a capable outline of the potential of knowledge level studies. Unfortunately, unlike Marr, the majority of practising cognitive scientists seem more concerned with finding actual process models of cognition than with exploring the scope and structure of those processes at the knowledge level first. They attempt to discover the specific representations and control mechanisms people rely on in performing their tasks, rather

than analysing task knowledge per se. The majority of research conducted on memory, reasoning and decision-making, for instance, involves constructing models in which the key hypotheses are precisely about the form of internal representations. Psychologists set out to discover whether perceptual memories are encoded propositionally or imagistically[1]; whether people reason with Venn diagrams[2], with notational variants of the predicate calculus, or in natural language; whether errors in syllogistic reasoning are the result of mistakes occurring during the encoding process -- reasoning proceeding afterwards with flawless logic[3] -- or whether errors arise through misapplication of rules of inference. The list may be extended. In each case the psychologist does not ask the abstract question: what knowledge is necessary or sufficient for task competence? But rather the specific question: what representation does the agent use? Knowledge, if a concern of these theorists at all, is inferred from representation and not vice versa.

Despite this trend there is an evident danger in moving to step two before completing enough work on step one: choices about representational format and algorithm may be made before enough facts

---

1 S.M. Kosslyn, "Imagery, Propositions and the Form of Internal Representations", reprinted in Ned Block(ed) Readings in Philosophy of Psychology, Vol.2, (London, Methuen, 1981). See alse Zenon Pylyshyn "What the mind's eye tells the mind's brain: a critique of mental imagery." Psychological Bulletin, 1973, 80 pp1-24.

2 Erickson, J.R. "Models of Formal Reasoning", in R.Revlin and Meyer, Human Reasoning, op cit

3 Martin D.S.Braine "On the Relation Between the Natural Logic of Reasoning and Standard Logic", Psychological Review, 85, 1978 pp1-21. See also R.Revlin and V.O.Leirer "The effect of personal biases on syllogistic reasoning: Rational decisions from personalized representations" in Revlin and Meyer, Human Reasoning op cit.

are in. Pat Hayes stated the danger in these words:

> It is perilously easy to conclude that, because one has a program which works (in some sense), its representation of its knowledge must be more or less correct (in some sense). Regrettably, the little compromises and simplifications needed in order to get the program to work in a reasonable space or in a reasonable time, can often make the representation even less satisfactory than it might have been....
>
> I emphasize this point because there is a prevailing attitude in AI [and to a somewhat lesser extent in cognitive science] that research which does not result fairly quickly in a working program of some kind is somehow useless, or at least highly suspicious. This may be partly to blame for the dearth of really serious efforts in the representational direction, and the proliferation of programs and techniques which work well (or sometimes badly) in trivially small domains, but which are wholly limited by scale factors and which therefore tell us nothing about thinking about realistically complicated worlds.[4]

Hayes was speaking about research in AI, but his comments apply equally to much of the work done in cognitive science. A look through the journals shows article after article devoted to explaining performance in highly specific tasks. As Newell noted[5], psychology, in its current style of operation, deals with phenomena which, once brought into existence by some clever experimental discovery, elicit a flurry of experiments to investigate it. As research, it seems driven by low-level discoveries of specific phenomena. Even such grand questions as whether people reason with Venn diagrams or in natural language tend to be assessed by reference to specific experimental tests. Perhaps this is inevitable: science grows from the specific to

---

4 "The Naive Physics Manifesto" appearing in D.Michie(ed). Expert Systems in the Micro Electronic Age, (Edinburgh: Edinburgh University Press, 1979).

5 "You Cant't Play 20 Questions With Nature and Win" in W.G.Chase (Ed.) Visual Information Processing (New York: Academic Press, 1973).

the general. Or again perhaps there is no other available means of proving a general claim than by presenting what one thinks is 'the crucial experiment'. But all too often models are constructed whose application is specific to narrow tasks. The result is a plethora of models and partial theories which bear no clear relation to others in the field save a general acceptance of the computational metaphor. As new models are postulated old models tend to be supplanted outright, leaving little explantory residue to be incorporated in future theory.

Despite its promise of a more structured or informed approach to the questions of cognitive science there remains widespread disagreement about the legitimacy of the knowledge level. Intentional states hold an ambiguous place in modern science largely because there remains doubt about the legitimacy of both the method of ascribing intentional states and the explanations in which they figure. Representations are a fit topic of research because they are thought of as formal entities. There is nothing suspicious about computational explanations, for they are as mechanistic as one could hope. States of implicit knowledge, on the other hand, have a disturbing lack of robustness. And it is not clear what sort of explanation one has when a knowledge theory has been provided. Why bother with knowledge when one can have representation?

Before I try to allay some of the worries about the methodology of knowledge level research I want to offer three reasons for bothering with knowledge in the first place. Thus far I have argued that knowledge may be a guide to mechanism; that without analyses at the knowledge level representational studies will be blind and ad hoc. We can, however, be more explicit.

The first reason has been eloquently defended by David Marr:[6] knowledge level research may provide a level of analysis where discoveries and theories have a lasting value. Specific theories of representations come and go; sometimes they are rejected because of some dispute over the information -- the knowledge -- that is to be represented. But often the reason one is preferred to the other is simply that it provides a better account of how information is represented, where 'better' means 'is more simple or elegant or natural given our current ideas about the human cognitive processor'. It is important to distinguish these reasons. Discoveries about the human cognitive processor can always shed light on the likelihood of certain representational formalisms. As we discover new facts about the processor new representational schemes will seem more natural. Old ones which once seemed natural will be outdated. By contrast, at the knowledge level once a genuine cognitive task has been identified and a theory provided for it, no discovery about our functional architecture can weaken the theory. Some doubt may remain about whether the task has been correctly specified. If it has though, then the results derived at the knowledge level hold up, and need never be done again. In this respect they are like results in mathematics: once knowledge is shown necessary or sufficient for a task it remains so.

Second, the very act of thinking about a task as a knowledge-involving project may help us to understand the task better. One problem of cognitive science is that we are unclear about the natural tasks which the human mind has been designed to perform. If we were

---

6 "Artificial Intelligence: A Personal View", Artificial Intelligence, 9, pp7-48.

quite clear just what the basic cognitive tasks are, we might be able to devise a better class of experiments to discover their nature. For a while psychologists defined cognitive tasks in terms of the behavioural tests that were meant to display mental abilities. Things are slowly changing. Certain cognitive tasks may now be viewed as internal tasks as, say, achieving a certain knowledge state. Even if we do not go this far, and we continue to individuate mental abilities by reference to external tasks, there remains a serious problem in grouping those external tasks into natural classes. How can we put boundaries on the range of tasks in which a person may display his semantic competence, for example? Reading, writing and speaking involve semantic capabilities, no doubt, but semantic processing may take place during perceptual tasks, or in reasoning, or in... . By thinking of behaviour and behavioural tasks in terms of the sort of knowledge which could be involved in performing those tasks we may begin to identify which cognitive abilities are being exercised when. Hence we may be able to posit ways of modularizing or dissecting complex abilities and measuring the contribution of each component to the performance of a given behaviour task.

Nowhere is this prospect exploited more than in linguistics. The modularity of knowledge is one of the central claims made in modern transformational grammer and is driven as much by considerations of natural distinction at the knowledge level as by specific discoveries of the mechanisms of speech. Chomsky's avowed programme, for example, is to discover the way competent speakers organize their knowledge of their language and to show the natural categories of knowledge involved in speech and the way they are grasped and systematically deployed in

the speech 'system'. He conceives of the mind as a system of mental organs, each organ having its own specific structure and function. The way to discover the organs is by analysing intuitions we have about natural divisions at the behavioural and knowledge level simultaneously.

For instance, students of language are strongly motivated to draw a distinction of kind between lexical knowledge -- so called analytic knowledge -- and empirical or synthetic knowledge. As George Miller once wrote

> The advantages of distinguishing lexical from practical (empirical) knowledge is that it helps to set manageable bounds on what phenomena a theory of linguistic communication can be expected to treat. There is more than enough even if we limit our theorizing to the lexical meanings of words and phrases and their linguistic entailments. If we must also include a theory of knowledge in general the theoretical task will become unmanageable. Moreover, the difference is readily illustrated. One feels that the relation of 'John's children are asleep' to 'John has children' is very different from its relation to 'John is married'. In the latter case, an inference can be justified by practical knowledge, but it lacks the requiredness of the former relation, which can be justified in terms of the linguistic meaning of John's children, in terms of lexical and linguistic knowledge.[7]

Miller's argument rests on considerations about the behavioural scope of linguistic theory and on considerations about the reality of the distinction between lexical and practical knowledge. Philosophers have long claimed there is a distinction in principle between these two forms of knowledge. The two (arguably) are acquired in different

---

7 "Practical and Lexical Knowledge" in F.Rosch and B.B.Lloyd (eds) Cognition and Categorization (Hillsdale, N.J.: Lawrence Erlbaum, 1978). p 306.

ways: lexical knowledge through an appreciation of when one is entitled to use a sentence solely on the basis of sentences already asserted, and practical or empirical knowledge through an appreciation of the sensitivity of certain sentences to the non-linguistic context in which they are uttered. Moreover, it is arguable that they have a different epistemological status: analytic inferences are true a priori, if true necessarily true; synthetic inferences are true a posteriori, if true contingently so[8]. If the philosophical distinction has a grounding in the way our minds work it suggests that lexical and empirical knowledge may be stored in different places, or have natural connections with different modules. This is not to say empirical knowledge is not constantly called on in the course of semantic processing. But (arguably) it is used in a different way, and enters in a different time and place in the processing path. The distinction also plays out at the behavioural level. Our practices of justification mirror our awareness of when inferences are analytic and when synthetic. We therefore intuitively distinguish tasks at the knowledge level in a way that mirrors our distinctions at the behavioural.

Whether or not Miller is right in his claim about the separability of lexical and practical knowledge, his method lies squarely at the knowledge level. By re-orientating our thinking from the behavioural to the knowledge level, we can speculate about the inner machinery of

---

8 It is irrelevant for my point that these classical philosophical distinctions have been challenged by philosophers such as Saul Kripke in "Naming and Necessity", in Davidson and Harman (eds) Semantics of Natural Language. Distinctions are always challenged. What matters in empirical research is whether such classifications are serviceable.

the mind that is responsible for complex behavioural displays. If we think of cognitive competence exclusively in behavioural terms, we have the classical problem of operationalism: how do we decide when performance on new tests is a manifestation of the same ability? By appealing to natural categories of knowledge, (if these can be found and justified) we have a theoretical position from which to classify behaviour and behavioural tasks. The very act of thinking about a task as knowledge-involving helps us to ask the right questions about the task.

The final reason for thinking of mental competences in terms of bodies of knowledge is that it may make the problem of describing and explaining learning easier. One of the legacies of behaviourism is that when a psychologist thinks of the changes a person undergoes when she learns how to perform a given cognitive task he thinks of the changes in her behavioural dispositions. What kind of order can he hope to find in these changes in behavioural dispositions? As a person learns more about her world she changes her behaviour and dispositions in subtle and highly complex ways. There may be no apparent order in the way dispositions change in virtue of learning a few new facts. Undoubtably learning must be capable of being manifested somehow in behaviour. But the manner in which it it will be manifested depends on other factors -- among other things, on other knowledge and desire states the agent has. Thus the results of learning do not cash out directly in behavioural change; they are mediated by other internal states. Since the interaction of these other states is potentially complex, the impact of learning is not easy to define at the behavioural level. Hence, although we may describe learning in a

simple way at the knowledge level, we cannot describe it in any simple away at the behavioural level. The behavioural changes associated with learning may be virtually invisible. The only way to recognise these behavioural changes is as changes in the agent's knowledge. Moreover, since the cognitive theories we accept should describe representational mechanisms that can be learned it is probable that additional study at the knowledge level will help us to define dynamic constraints -- learning constraints -- which those mechanisms must meet; they must be able to change through time appropriately.

I think these three reasons constitute a strong motive for developing a research program at the knowledge level. If we can legitimately attribute knowledge to a system, and then reason about the logical relations which that knowledge bears to the task -- for instance, whether the knowledge is necessary but not sufficient for the task, or sufficient but not necessary -- we can begin to lay down a body of theory which may be called on to help understand the design of task competent systems. This will be useful whether we intend to design systems ourselves, as in Artificial Intelligence, or we intend to discover the design of systems already existing, as in cognitive psychology. The chief impediment is that concepts such as task environment, task knowledge and rationality, the corner stones of knowledge level research, are still largely unanalysed. Newell and Simon, for example, have used the concept of task environment extensively in their research on human problem solving,[9] but they attempt to explicate the concept primarily through examples. Moreover,

---

9 Human Problem Solving, (Englewood Cliffs N.J.:Prentice Hall,1972).

they have restricted their examples to problems that are nicely structured and fairly well understood at the behavioural level. We may not know how a chess player figures out his next move, but we have a good idea of the behavioural displays that constitute chess play.

Where our concern is with tasks that may take place inside the head, however, tasks which may not be correlatable with a distinct set of behavioural performances, we will have to broaden our conception of task environment if we are to see these inner tasks as genuine tasks. In the third section of this chapter I look at the background of the task environment concept in an effort to extract its main properties. These properties are sufficiently abstract to allow us to extend the notion to tasks which, in one sense, have no 'behavioural' environment at all. Of course, their environment exists; it is just abstract, and the constraints on its structure come from logical features of the task. I then consider how this abstract notion of a task environment may be applied to the study of semantic processing. I conclude with an examination of the meaning of necessity and sufficiency analyses at the knowledge level. David Marr's work figures heavily in the section on necessity.

To begin though, let us turn our attention to the process of knowledge attribution. The problems and promises of knowledge level research all emanate from this basic process.

Knowledge Attribution

At the knowledge level a system is seen as having three fundamental cognitive parts: a knowledge base, a system of desires and rational faculties endowing it with the capacity to harness knowledge in the

service of desires. See Figure 2.1.

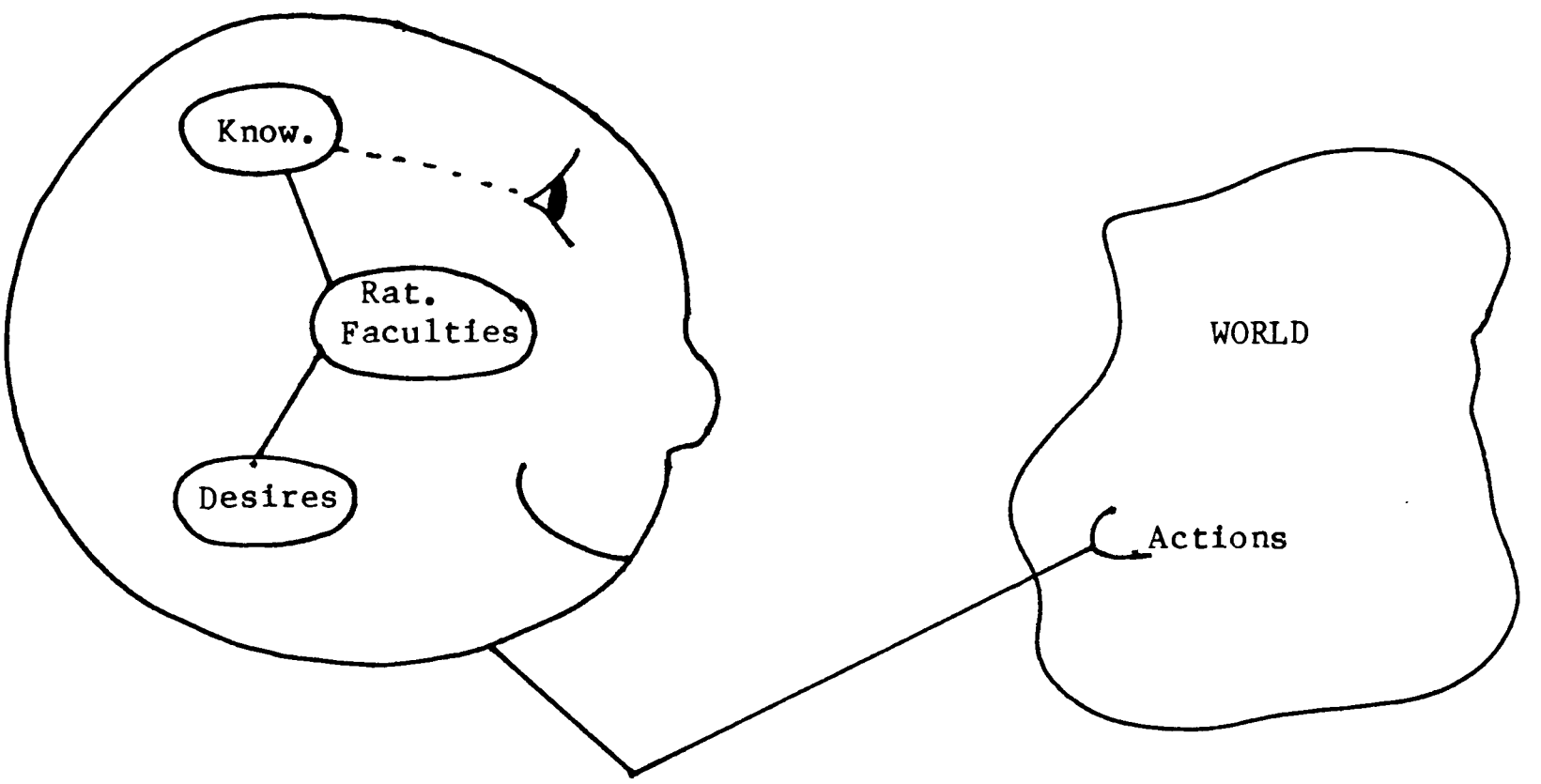


Figure 2.1

If we identify rational faculties with some function **rat** , knowledge with a set of propositions **K** , desires with a set D and actions with **A** ,then we may say that $A = \mathbf{rat}(K,D)$. Knowledge attribution is the process of extracting K from a description of the agent's desires, actions and rationality function. It is the process of computing the inverse function $K = \mathbf{rat}^{-1}(A,D)$.

Because we can decide what an agent knows only if we already know both what he desires and what his rationality function is, the knowledge attribution process begins by assuming we can recognise what the agent is trying to accomplish, by assuming that we can identify the rational point of his actions. This strong assumption is justifiable only if we may assume that as humans we share and can recognize a common world of tasks, social demands and problems, and share a common form of rationality.

The modern method of knowledge attribution actually began with Max Weber in the 19th century. Weber[10] argued that the social world is made up of a constellation of typical tasks, or social situations, social roles, each with its own structure, and each making its own characteristic demands on agents. Agents have desires appropriate to their tasks, and we can recognize their tasks because we share a general world-view. Within this world view, human life can be represented as a passage from task to task, environment to environment, and human agents can be represented as collections of task competent beings. Thus one and the same person could be carpenter, mayor, English speaker, cricketer, and father, as well as an agent capable of performing sums, driving an automobile and so on. Research in the social sciences was to be carried out by breaking up the social world into distinct task spaces -- into social niches as it were -- in which agents, viewed as ideally competent performers, would act out their parts as determined by the exigencies of their situation. As long as each agent knew the rules of rationality (i.e. the rules of appropriate action) governing his behaviour in that space, he could move through life in a meaningful way, acting as we expect mature human beings to act. To use a modern phrase he would be well adapted to each of his niches.

---

10 See, for example, "Objectivity in Social Science and Social Policy" and "Basic Sociological Terms", both reprinted in Fred Dallmayr and Thomas A. McCarthy, Understanding and Social Inquiry, (Notre Dame: University of Notre Dame, 1977); and articles by Alfred Schutz, such as "Concept and Theory Formation in the Social Sciences" in Dallmayr ibid; and "The Social World and the Theory of Social Action", "The Dimensions of the Social World", and "The Problem of Rationality in the Social World", all reprinted in his Collected Papers Vol.II (The Hague: Martinus Nijhoff, 1976).

Weber's approach defined three questions for the social scientist to answer:

(1) How is the social world partitioned into task spaces? What are the 'natural' tasks of social life?

(2) How is each task structured? What are the objective features of the different tasks (social roles) which determine how an ideally rational agent will act?

(3) What are the rules of rationality associated with each task or social role?

Working within this paradigm the process of knowledge attribution unfolds in four stages. We begin by ascertaining that we are dealing with a rational system, a system much like us in having intelligence, learning capacity, and diverse skills. We then isolate the different tasks or social niches it is adapted to and single out one task, for particular study. Then we analyse the task environment. Finally, we compare the performance of the system to what we calculate would be optimal performance in that task; if the system performs ideally it has perfect knowledge of its task environment; if its performs sub-optimally it has less than perfect knowledge. Thus the entire attribution process turns on being able to analyse the task environment and being able to decide what would count as optimal conduct.

This approach is an instance of what may be called the adaptationist strategy.[11] Each agent is assumed to be more or less well adapted to his social niche. Hence his conduct can be expected to approximate some ideal, some optimum that is definable by studying the

---

11 The term is Simon's. See, for instance, his essays "Cognitive Science: The Newest Science of the Artificial" Cognitive Science,1, 1980; and "Theories of Bounded Rationality" chapter 8.

"objective" conditions of the niche. We can infer knowledge because the system must know enough about its task and situation to figure out how to act optimally or near optimally given his desire to accomplish his task. Deviation from the ideal is proof of the absence of knowledge or inadequecy of rational faculties.

Weber's concern was with the partitioning of the social world. He was a sociologist above all and interested in the structure of social space and agents' inner knowledge of that social space. Human action could be explained, he maintained, by appreciating the motives people have and their knowledge and beliefs. The adaptationist approach was a way of gaining insight into their mental states by assuming that society has a set of niches, which force people to play certain roles. By studying what an agent ideally suited to a niche would have to want and know to be able to act out his part perfectly, we could discover the probable knowledge and desires of people actually playing those social roles. The basis for knowledge attribution was laid.

In cognitive science our concern with the social world is not paramount. The tasks and environmental demands we are interested in exploring are not particularly social: they are cognitive. We are interested in tasks that reveal the natural partitioning of our cognitive talents. Hence the partitioning of tasks we are interested in, is not a partitioning of social space into social problems, social roles and so on; it is a partitioning of cognitive space, the world of basic intelligent functions. We want to know what are our major cognitive rôles: should we divide ourselves up into speaking man, thinking man, seeing man, moving man, and so on? Is speaking man really a composite of other cognitive agents -- a syntactic agent,

semantic agent and pragmatic agent? These are the sorts of questions cognitive scientists must ask. They must

(1) partition cognitive life into a system of 'natural' tasks;

(2) state how each task is structured; and

(3) state the rules of rationality appropriate to each task.

Once this is done they will be in a position to evaluate the necessity and sufficiency of a given body of knowledge for a task. And they will be able to speculate on the representations and mechanisms instantiating some of those bodies of knowledge. But all such possibilities depend on having a clear notion of a cognitive task which the agent performs. It is to that notion I now turn.

The Task environment

A task environment is an amalgam of two things: a task to be performed, and an environment of action in which to perform it. The knowledge a rational agent would find sufficient or necessary for a task depends on the possibilities, consequences and potential interferences he will encounter in the environments in which he will be operating. If he can reasonably expect outside influences interfering in his normal activities -- if, for instance, he can expect disruptions in his job from telephone calls, equipment failure, or secretarial inefficiency -- he will have to know enough about the nature of these interferences to formulate strategies that can accomodate their effects. Planning involves knowledge of contigencies. This applies whether the task is a 'social' one or a 'cognitive' one.

The environment that is relevant to planning cannot be identified with some concrete fragment of space-time. What is relevant in

planning is possibilities; potential consequences of possible actions, potential interferences, potential constraints and so forth. It is this abstract construct, the union of possible task relevant features, that is the task environment.

Construed in this abstract way the concept of a task environment can be applied to any definable task. It does not matter whether the task is performed in physical space by manipulating objects or by moving one's body, in mathematical space by manipulating symbols of numbers, or in the mind by manipulating images, thoughts or other representations. What counts is that there is a goal and a set of actions that may advance the agent toward that goal.

The origins of the concept grew out of reflections on simple behavioural experiments. In classical learning experiments, for instance, a rat or other laboratory animal is placed in an experimental set-up, such as a maze, it is made aware of a potential reward, and encouraged to make 'choices' at stimulus points marked by the experimentor. From the psychologist's perspective the environment is partitioned into a universe of choice points, each stimulus representing a particular range of possible behaviours or choices the organism can undertake. Some choices bring it closer to its goal, others take it further. The objective of research is to fit the animal's performance on repeated trials into a pattern, to see how it learns or adapts to its task environment.

The task environment is not simply the maze, but the properties of the maze that may bear on the rat's attainment of its goal. Properties must be task relevant. Behaviourists at one time tried to minimize the place of goals in animal behaviour claiming to be looking for habits

that arise through selective reinforcement. But in any analysis of behaviour some antecedent restrictions must be placed on the sorts of behaviour that the experimentalist will consider as suitable for reinforcement. The tacit assumption has always been that only potentially goal-relevant actions will be considered. Thus when a rat is placed at the crossroads of a maze and food is placed at the end of one of the paths, the only actions the analyst will consider the animal to have available for reinforcement are movements along a path. Scratching, salivating, sniffing, no less than breathing and perspiring, are not considered actions that are rewarded in mazes. Why? Surely not because they belong to different categories of behaviour: rates of breathing, sniffing, salivating, in certain experiments, can be the operants conditioned; it is because they are not relevant to the goal in question; they are not moves that can advance the animal toward attaining its objectives.

Intrinsic to the concept of task environment, then, is the concept of a goal the subject is trying to achieve. This is a point of some importance. When Weber first introduced ideal type sociology his motive was to avoid having to discover the actual goals of each agent by assuming that if the agent was found playing a certain social role he necessarily had the appropriate goal structure for the role. There were limits to this approach: ideal types had ideal goals, real types tend to fall short of the ideal. Nonetheless, deviation from the ideal could be accomodated, Weber thought, as long as there was a systematic way in which actual agents depart from the ideal (biasing). The process of attribution and explanation, consequently, began with analysis of perfectly rational conduct. Human inadequacies would stand

out against this background of ideal action. Thus the way to attribute mental state was through analysis of the structure of the task environment itself, for only relative to a well defined task can we define ideally rational conduct.

Knowledge attribution goes hand in hand with task analysis. What then are the major constituents of a task environment? If we generalize the notion of task environment used in our dicussion of the maze model we can distinguish four paramenters in terms of which we describe and classify task environments.

1) a repertoire of possible goal-relevant forms of behaviour;

2) a network of choice points, (the crossroads of the maze);

3) a function mapping behavioural options onto choice points (the paths open); and

4) a consequence function determining the goal-relevant effects, or probable effects, of an action at a choice point (the consequnces of following a path).

A change in any one of these parameters would constitute a change in the task environment.

Defined in this manner task environments have no special connection with space. This is desireable because many cognitive tasks may be internal to the mind; they do not involve behaviour in a physical environment. The parameters of these tasks, therefore, will not be found by studying the physical environment of the creature. In many cognitive tasks neither the behavioural repertoire, the network of choice points, option function and consequence function can be read off from the environment by studying its physical structure. Nonetheless since we are beings living in the physical world many of the tasks we

routinely perform do occur in physical space-time. To return to the maze example, the choices facing a rat at any choice point are given by the physical paths available. It may move forward, backward, to the right, to the left or stand still. Usually it must choose between all five options, though toward the sides, choice is restricted by boundaries. Nowhere will the experimenter, or the rat for that matter, have difficulty in reading choice points and options from the physical structure of the maze. The same can be said for the consequence function, for as the topology of the maze is a physical topology the eventual consequence of following any path can be discovered just by studying the physical environment.

Again, when we assemble a structure, a car engine for example, the task unfolds in space and time. The environment now is defined by the parts available; and the way they fit together. The task is to find a sequence of behavioural manipulations of parts that configures the engine correctly. When humans attempt either of these tasks -- to navigate in a maze, or assemble a structure -- there is evidence suggesting that they create an internal representation of the task environment in which to monitor their actions and evaluate possible paths (sequence of choices). The adequacy of those internal representations is obviously a function of their structural resemblance to the task environments they represent. The more structural features that are internally represented, the more it is probable that an agent's internal representations will be effective in planning.

It is worth distinguishing mechanical constraints on task structure from logical constraints. Given a world like ours the laws of nature impose definite mechanical constraints on the paths that are physically

possible. For instance, in automotive assembly the sequence of assemblies we could try is restricted by the parts and the laws of physics (assuming we are adequately strong and dextrous to manipulate the parts as necessary). The class of actions available at any choice point depends on the parts that remain, the current structure of the engine, and the laws of nature. We cannot change at will the order in which the engine is put together: certain precedent relations must be observed. Objects cannot be taken out of closed containers or new ones inserted. That is a physical law pertaining to rigid bodies. Consequently, the crankshaft must be in place before the piston can be inserted. Similarly, we are obliged to build the inside of things before we build and seal their outsides. Again, it is a law of nature that solid objects cannot simultaneously occupy the same space. We cannot substitute new parts for old without removing the old.

Logical features of the task, by contrast, flow from the very definition of the task. Thus it is a logical feature of assembling an engine that the finished product will have more parts than it had when entering the assembly phase. Or again, that each part will be connected to the engine. In games like noughts and crosses it is a logical feature of the task environment that one cannot get three in a row before getting two in a row and before placing one in a row. The laws of geometry plus the rules of the game determine these precedent relations. They hold whether the game is played with solid pieces in worlds like ours, or with symbolic pieces in imaginary worlds with different physical laws. The constraints flow from the task itself.

I have focussed on these different features of task environments because analyses of task environments are at bottom analyses of

structural, i.e., logical and mechanical, constraints on task performance. These are the cardinal properies agents must in some sense know. Structural constraints determine the option function since they define the actions open at all the choice points. They determine choice network since they indicate when a choice is available. They determine behavioural repetoire since an agent's possible acts in an environment are determined by the totality of options at all choice points. And they determine which environmental states are accessible from others. Graphically this structure may be represented as a maze or tree, but there is nothing priviliged in this representation. At times it may be more convenient to state the constraints explicitly as laws defining state change, etc.

Natural Tasks

So far our analysis of the environment of the task, has been uncomplicated by questions of its naturalness. We regarded agents to be well adapted to their task environment but left it an open question whether those environments are characteristic of the sort evolution may have favoured. This is reasonable if we assume man's cognitive system has the power of a universal Turing machine and can in principle be programmed to act as if it is adapted to any cognitive environment. Even universal machines, when instantiated in hardware, however, have specific functional architectures, the considerations most relevant to designers concern the efficiency of the machine with respect to its natural tasks. It was mentioned that the tasks of interest for cognitive science tend to be abstract. In principle this poses no special problem for analysis. It hardly matters whether the 'laws'

regulating task performance come from physical theory, mathematics, or sociology. Each defines a space of possible actions in which performance can be interpreted as the selection of choices within the space. But from the standpoint of design not all tasks are equal; some are more informative about the structure of the basic cognitive system.

The tasks cognitive scientists study are not tasks people normally undertake in the course of their daily lives; they are set up experimentally. Cognitive science is not cognitive ethology; it is an experimental science in which performance in an artificially created situation is observed. Accordingly it suffers the same drawback all non-ethological psychology suffers: we cannot tell whether the task environments experimentally created reflect tasks which are ecologically natural. This worry is exacerbated by the standard practice in cognitive science of creating tasks where agents' performance systematically deviates from the perfectly adaptive, suggesting that the task is not one evolution prepared us for.

The complaint is one ethologists have been lodging against experimentalists for some time now. As Tinbergen said in 1957:

> Ethologists tend to spend much time in preparing full and accurate descriptions of the whole behaviour pattern of a selected species, as preliminaries to further study, whereas behaviourism has for the time being abandoned this task, and concentrated on detailed analysis of the causation of selected simple units of behaviour.[12]

Taken in isolation from its normal context, it is unclear what salivating to the sound of a tone or pressing a lever to obtain food in an operant apparatus teaches us about an organism. As we discover more

---

12 Preface to Instinctive Behaviour, (Ed.) Claire Schiller, (London: Methuen, 1957) pp.xvi-xvii.

about innate constraints on learning and about how organisms naturally partition their world into choices, we are forced to rethink the significance of behaviourist experiments. Even today, when behaviourism is unfashionable, experimentalists often go directly to the laboratory without researching the work of ethologists to consider how behaviour might be ethologically segmented, how what seems disconnected or pointless from an experimental perspective fits into a functionally significant pattern when viewed ethologically.

In ethology integration of behaviour is essential. As Hinde has remarked:

> In studying the causation of behaviour, ethologists recognise the need not only to analyse the causal bases of particular aspects of behaviour, but also to re-synthesize the products of analysis in order to understand how functional systems of behaviour are interpreted.[13]

We know that what appears disconnected and 'task-specific' from an experimental prespective may be under the control of larger more powerful control units. Behavioural fragments may be part of a behavioural system which unifies apparently diverse behavioural performances. Experimentally it may be hard to discover those natural suture lines of behaviour.

In this section I want to consider whether studying systems at the knowledge level can help us to isolate and define our fundamental cognitive faculties and our fundamental cognitive activities. If we view cognitive faculties as storehouses of knowledge, one differentiated from another by its content, might we be able to map out

---

13 Ethology, (Glasgow: Fontana, 1982), p.95.

our gross cognitive architecture? A plausible method might run like this:

i) consider faculties which are absolutely central to human life: seeing, speaking, reasoning, evaluating;

ii) identify clear cases where these faculties are displayed in behavioural tasks;

iii) abstract a core of knowledge involved in those behavioural tasks;

iv) recast the objective of the faculty as attainment of certain occurrent states of knowledge;

v) study the task structure of this newly defined faculty; and

vi) infer possible behavioural applications of the occurrent knowledge generated.

Step (iv) is the key to the method I am proposing. If we can recharacterize cognition non-behaviourally, seeing it as an activity entirely in the agent's head, aimed at the achievement of certain occurrent states of knowledge, we may be able to divide the cognitive system into a set of key components, which interactively are responsible for entire behavioural systems.

The idea is not new. In Rules and Representations, Chomsky has argued forcefully for giving up behavioural characterisation as the task of cognitive science. But the twist I am suggesting is that we also reconstrue the task environment of our various cognitive faculties. The way to motivate seeing our cognitive faculties as bodies of knowledge is to see their output to be occurrent knowledge states. Linguistic understanding, for instance, is not like running through a maze or putting a car engine together, both processes that

unfold in space-time. It occurs inside the mind and the achievement is occurrent knowledge. The advantage of so specifying it is that we are not committed to a set of necessary behavioural manifestations. Mediating linguistic behaviour are other mental states and dispositions. The range of ways a person might manifest knowledge of a sentence's meaning is astronomical. He might assent to the sentence if he believes it and is asked about it, he might refrain from answering questions about it but adjust his future non-verbal conduct to accomodate its truth, or he may simply ignore it. In general, there is no limit to the appropriate or rational responses he might make; semantic competence can be manifested endlessly in behaviour. Thus if we do not specify the primary or canonical manifestation of semantic competence in terms of occurrent knowledge states we shall have to identify its canonical manifestations with extremely complex dispositional states which we either know too little about to specify precisely or which we identify with behavioural abilities which in all probability artificially confine the scope of undertanding. In either case we cannot hope to correctly analyse what we achieve when we understand a sentence.

The attraction of interpreting certain cognitive abilities as directed at achieving occurrent knowledge states, then, is that we can finesse the problem of full behavioural specification. We need not state the diverse ways task competence can be manifested behaviourally because it is of the essence of knowledge states that their influence on behaviour be mediated by other knowledge, belief and desire states. We can calculate behavioural effects only if we assume the agent has certain other beliefs and desires. Consequently the objective of

cognitive research is to understand the structure of the processes leading to occurrent knowledge. It is to discover the basic structure underpinning cognitive performance; that is, the basic units and operations that are found in the paths leading to understanding, to sight or to whatever achievement is being interpreted at the knowledge level.

Let me explain this idea further by explaining how we should interpret the task of semantic processing.

The Task of the Semantic Processor

Let us assume that the goal of semantic processing is to produce occurrent knowledge of the meaning of each well-formed sentence we hear. From a string of phonemes, for example, the semantic processor must generate as output, knowledge of the meaning of that string. See Fig. 2.2.

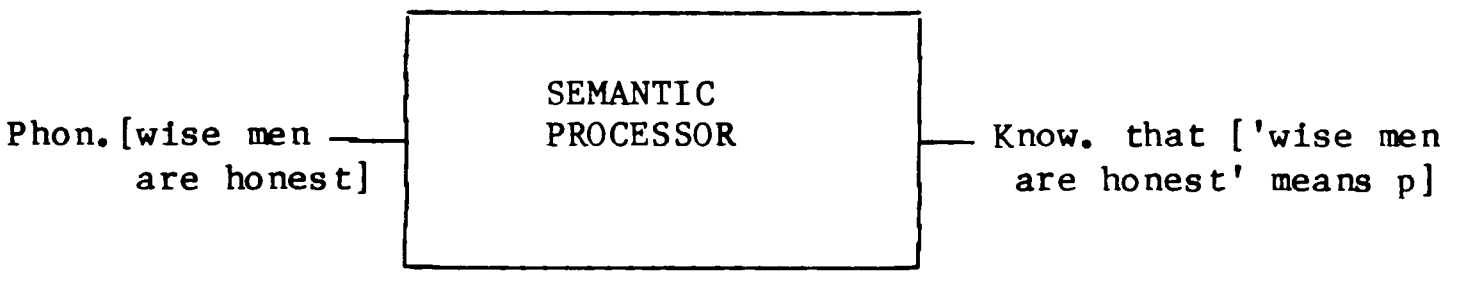


Figure 2.2

To analyse the task of the semantic processor we must discover the basic features or properties of its environment and the basic units and operations involved in performing the task.

As should be obvious from our discussion earlier, unless the goal of the task is well-specified it is impossible to state the logical or

mechanical constraints that might restrict the paths that lead from input to knowledge. We must specify exactly what knowledge of meaning consists in. But here we have a problem. For having given up the idea of identifying the goal of semantic processing with successful performance on some behavioural task(s) we are obliged to formulate a more abstract representation of its goal. As condition ii) implies, this goal state, the state of knowing the meaning of the sentence, must be such that it allows us to rationally explain obvious displays of semantic knowledge. To be more specific, occurrent knowledge of meaning should allow us to explain: (a) why and how we can recognize semantic relations between sentences when they are presented to us for comparison; (b) why and how we can explain the meaning of sentences to others, for instance, by paraphrasing the sentence, or describing when the sentence can be used, or the conditions under which it would be true; (c) why and how we can state entailments and presuppositions of the sentence; and (d) why and how we can recognize whether currently observable conditions satisfy the sentence. All these abilities are thought to flow from our occurrent knowledge of meaning. To be sure, other cognitive faculties are involved. We must be able to perceive, to speak or write and so forth. We also must exploit non-occurrent semantic knowledge. But occurrent knowledge of meaning must be so structured as to fit together with the knowledge used in these other faculties to produce the hallmark displays.

Now if an adequate empirical defence can be provided for a given analysis of meaning, we can redefine the task environment of the semantic processor and begin to explore its new structure. This new task environment I shall call its natural environment . For instance,

according to standard truth conditional theories of meaning, linguistic understanding consists in occurrent knowledge of the truth conditions of the sentence heard. Generating meaning consequently is equivalent to generating truth conditions. Given (1) an ordered sequence of a sentence's semantically significant elements; and (2) the truth axioms and combinatorial rules that apply to those elements, the system should be able to derive the truth conditions for the sentence. Thus the natural task environment of semantic processing would be a constructive space, not unlike the space of engine construction, where the parts manipulated are the basic semantic constituents of language, the parts mirroring the constituents of facts or states of affairs[14], the basic stuff of the world.[15]

If the truth condition[16] approach is correct two points follow. First, since the task becomes a simple combinatorial task, it is structured by the precedent relations implicit in the combinatorial rules. For instance, if there is a combinatorial rule specifying that a sentence coupling a name with a predicate is true iff the object denoted by the name satisfies the predicate, then we must find the denotation of the name and the extension of the predicate before we can

---

14 Cf. Loar "Each state of affairs is a logical complex having as 'constituents' the entities relevant for the truth of any sentence describing it." p.243.

15 Cf. Wittgenstein "The world is the totality of facts not of things." Tractatus

16 Although I am assuming a truth conditional semantics for the example, all arguments and conclusions apply to any axiomatizable semantic theory, whether based on truth, warranted assertibility falsifiability or verification. [If such we had.]

discover the truth conditions of the whole sentence. Second, because the units, as defined in the semantic rules are themselves semantic -- for instance, representations of the objects the agent can refer to -- and because the input to the system is non-semantic (that is, it is phonetic representations), the system cannot actually begin the truth conditional phase of analysis until it has translated phonetic units to semantic units. Thus we must parse the sentence so that we can identify which phonemic unit corresponds to which semantic unit. (This complicates the natural task of semantic processing and may encourage us to divide the task into different natural tasks.)

It is easy to misunderstand the consequences of such an analysis. A common complaint against truth conditional semantics is that people do not really encode truth axioms. Nowhere in our minds do we have sentences written in an inner language of the form "X satisfies 'is a cat' iff X is a cat". Children can learn to speak a language before they acquire concepts of words as linguistic entities capable of referring to non-linguistic entities. They may already have such concepts in their inner mentalese, but the hypothesis is unattractive. How then could a truth conditional theory of meaning teach us anything about our mental processes? It's claims are all non-psychological.

I hope it is clear by now that to argue in this way is to misinterpret the point of task analysis. It may be granted that nowhere in our minds do we have an axiomatic theory of meaning. But from this nothing follows about whether we obey the precedent constraints implicit in our axiomatic theory of meaning. If, as Frege suggested, a system capable of understanding must know the meaning of the words involved, and both the way to state this constraint and the way it

plays out in semantics is to represent it in the axioms of a truth theory, then we must know what those axioms express. But we certainly do not have to use those axioms in some encoded form in our semantic processing. If what I have argued in Chapter One is correct, we may incorporate the knowledge procedurally, or have it built into our processor. There is no guarantee that our semantic knowledge is explicitly represented syntactically in mentalese, and not much point in speculating; the answer is simply not yet in. But if a truth conditional approach is correct, we can say right now that our semantic processes are constrained by the essential precedence relations implicit in a truth theory of our language. The key question, therefore, is whether the theory is correct.

Philosophers and linguists are by no means in agreement concerning the correctness of truth conditional theories of meaning. For one thing, familiar homophonic theories offer little chance of explaining how we can recognize all semantic relations. A homophonic theory will explain why we can recognize entailment and formal contradiction, but it will fail to explain our capacity to note paraphrase, presupposition and analyticity, except of course in the question begging way of stating that the agent can recognize these relations because he knows the meaning of the sentences. I think such potential inadequecies of the theory can generally be resolved by bolstering it with axioms analysing some of the lexical items found in the standard homophonic axioms. For instance, by replacing standard homophonic axioms such as

An object satisfies 'is a bachelor' iff it is a bachelor

by a dictionary axiom such as

An object satisfies 'is a bachelor' iff it is a man and is unmarried

we can explain many, perhaps all, of our abilities to recognise analytic relations. Whether a supplemented truth theory will explain our full recognitional abilities in this fashion is a question still in dispute.

A more devastating criticism of the truth conditional approach is that it requires the agent to know things which transcend his powers, that an agent could never, even in principle, be in a position to learn the truth conditions of certain sentences and so is falsely attributed as knowing those truth conditions when he understands the sentences. Although it is beyond the scope of this essay to explore this question, it is noteworthy that it is most naturally raised at the knowledge level. Once we have an idea of the knowledge generated by various cognitive faculties -- of the type of occurrent knowledge the senses provide, of the knowledge the planning faculty provides, or the knowledge produced by our general purpose reasoning faculty -- we are in a position to ask whether knowledge of these sorts could find their way into the semantic processor, or whether the requisite semantic knowledge could be rationally derived from them. Questions of learning are easily raised at the knowledge level. Psychologists or computational linguists who rush to the representational level to find structures and formalisms that might encode semantic knowledge, only defer the need to ask foundational questions about the sorts of semantic knowledge that a competent system requires. Since languages are learnable an effective way into the language problem is to consider the sorts of things that can be learned. Questions of the exchange of

knowledge between components can be asked independently of how they encode that knowledge, if they 'encode' it at all, and independently of the primary behavioural tasks each component is linked to. We can ascend to a level of abstraction and contemplate conditions of rational exchange, rational development and so on. But more on this later.

Marr: Knowledge of Necessary Conditions

So far my discussion of the analysis of natural environments has remained at a lofty level, more programmatic than manifest. But in David Marr's work on visual processing[17] enough of the details have been worked out to reveal how something akin to the methodology I have been proposing can be applied to basic cognitive tasks. Marr took as his subject visual processing, and he accepted as his basic questions: What is the basic information processing task the visual system performs? What are the constraints intrinsic to the task? How does the visual system actually process visual information coming to the retina?

The visual system is a useful place to begin the study of cognitive science. First, most higher life forms have some visual sense, so there is evidence that nature is amenable to information

---

17 In a set of important papers, David Marr revolutionized the study of vision by explicating and deploying a powerful new methodology which lays particular stress on high level specification. Most of his ideas have been summarized in his book Vision, (San Francisco: W.H. Freeman and Company, 1982). But among his papers see: "Early processing of visual information."(Phil.Trans.R.Soc.Lond.B 275, 483-524 1976); "Visual information processing: the structure and creation of visual representations." (Phil.Trans.R.Soc.Lond.B 290, 199-218 1980); D.Marr & H.K.Nishihara, "Representation and recognition of the spatial organization of three-dimensional shapes." Proc.R.Soc.Lond.B 200, 269-294 1978); D.Marr & T.Poggio, "From understanding computation to understanding neural circuitry." (Neurosciences Res.Prog.Bull. 15, 470-488 1977); D.Marr & T.Poggio, "A computational theory of human stereo vision." (Proc.R.Soc.Lond. B 204 301-328).

gathering through sight. Second, perceptual knowledge is obviously a key component of cognition: it feeds into the reasoning system, the problem solving system, and so in higher organisms, issues in occurrent knowledge. And third, it is probable that there are certain invariants in the visual process. Being an information processing task that lies close to physics, vision is strongly constrained by the laws and principles of optics, image resolution and so forth. There is a large body of knowledge about the task environment.

Prior to Marr, studies of vision were undertaken by teams of investigators working on separate and apparently unconnected topics. Neurophysiologists were busy exploring the behavior of neurons at successively deeper levels of the visual pathway, hoping to correlate separate visual functions or capacities with specific neural activity -- though they had no detailed theory of what these functions were or what the end product of sight was. Psychologists were working on linkages between memory and vision, looking for principles concerning how background conditions, attentional set, and interests could affect what was seen -- though again they did not know how to characterize the ingredients of what was seen, the contents of visual knowledge. Philosophers, meanwhile, were concerned with the epistemological validity of sight; but they too spent little time describing exactly what aspects of the world we might come to know through sight. It was Marr's signal contribution to recognize that although genuinely distinct, these orientations could be related if one had an adequate characterization of the task at the knowledge level.

Marr's new approach to vision begins with the idea that vision is the computation of a description of the world from an image of it; it

is a process by which a description, an internal representation of what there is in the environment, is built up from an intensity array of light. Visual processing issues in occurrent knowledge of what is where in visual space. Worries about consciousness and the felt experience of what it is like to see are ignored. They are replaced by two more well-defined problems, namely, to state:

1) exactly what characteristics of the world our visual systems discriminate, and

2) what assumptions, computations and tacit knowledge must our visual systems rely on to recover these characteristics.

The first problem is an exercise in pure specification. It requires that we state what the system is capable of doing -- in this case, the occurrent knowledge it is capable of generating. Merely by looking, we gain occurrent knowledge of a host of visual properties: motion, position, orientation, shape, size, surface texture, distance and more. The second problem raises questions of tacit knowledge (and computational procedure): it requires that we state what a system must tacitly 'know' about the world (and what computational procedures itis capable of implementing) to come to occurrent knowledge of the visual world.

This way of putting the matter is not Marr's. He made few explicit remarks about knowledge, tacit or occurrent. But his way of framing questions constantly alludes to knowlegde gained through sight. and to principles unconsciously known. For instance, he wrote:

> The question we have to ask is, What assumptions are reasonable to make -- that we unconsciously employ -- when we interpret silouettes ... as three dimensional shapes. [18]

---

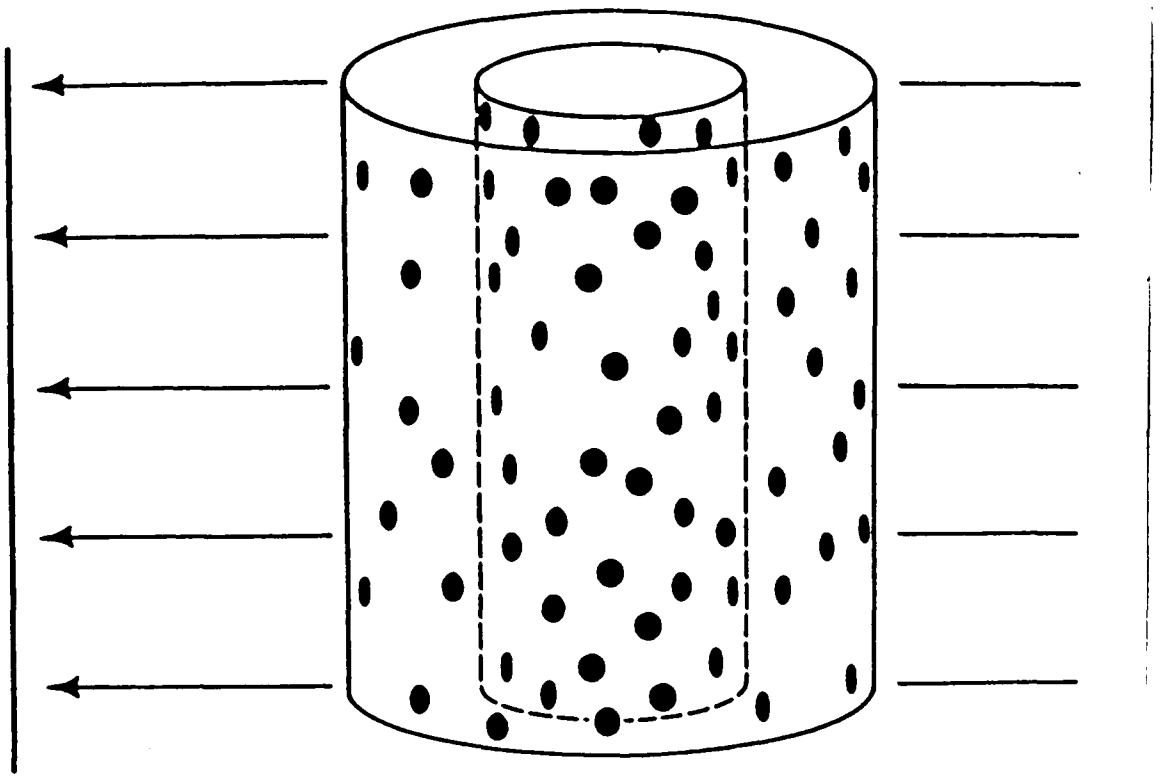
18 Vision p.219.

Any reference to unconscious employment suggests an analogy with knowledge level research. Marr's book Vision is an effort to answer both these questions. Obviously they are related. Only relative to an account of the input-output correlations which we know a system is capable of producing, can we talk of the knowledge a system must have in order to have those abilities. To prove A is necessary for B, one must first know what B is. But that is only a start. To prove necessity we must show that if A had not existed, if the system did not know X, then it could not have the ability in question. This is the heart of the necessity approach to the knowledge level and the essence of Marr's computational theories.

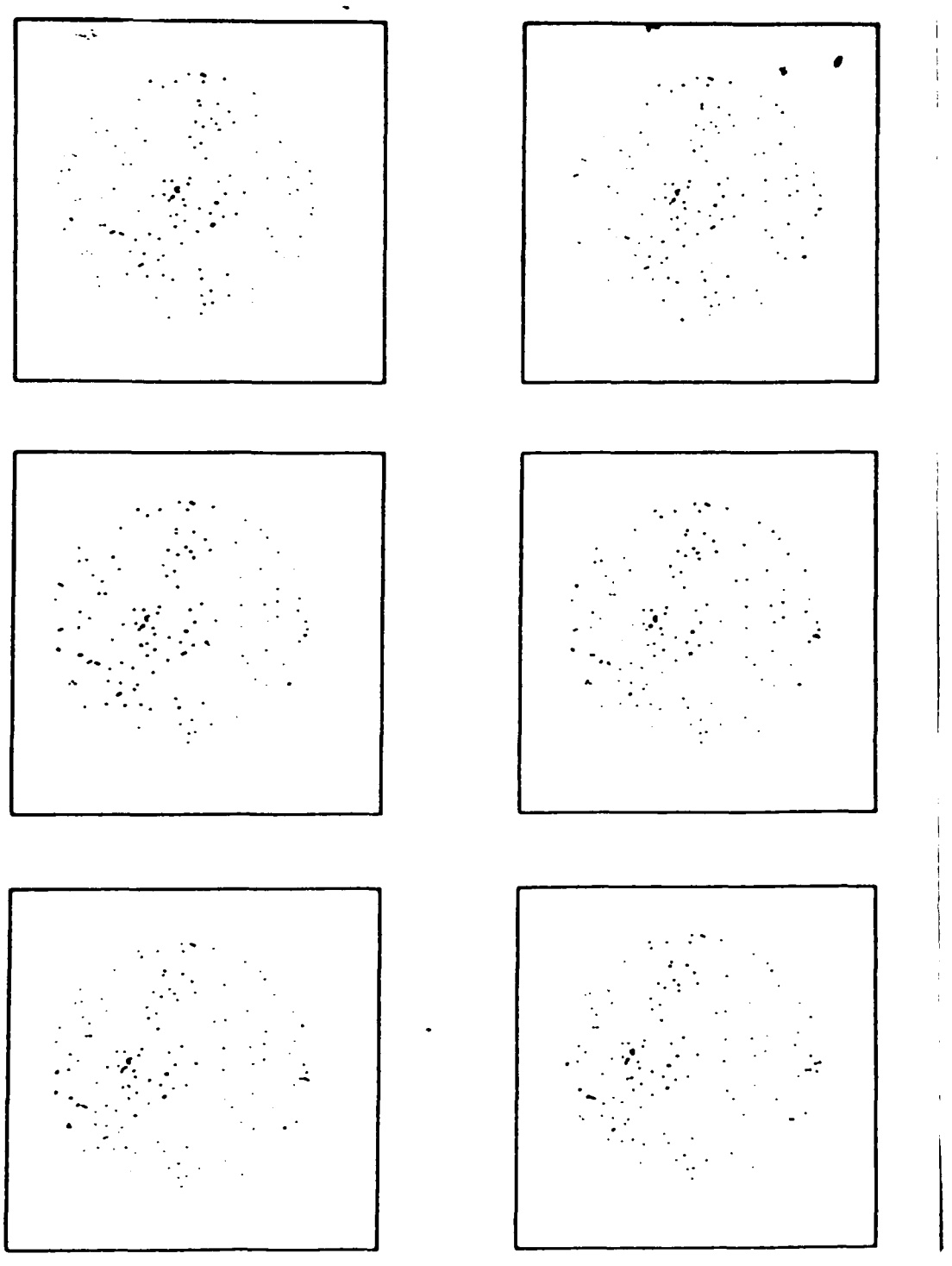
A clear example of such a necessity analysis is found in Marr and Ullman's[19] work on visual motion. It is known that limited information about the motion of points and shadows in the visual field can be used to make rather significant discoveries about surfaces and structures. In an elegant experiment Ullman established that if we are shown a flurry of dots that correspond to the projection of light through a network of holes in two revolving drums, (see figure 2.2) we can make sense of those projections as projections of the structures they really are, namely of two revolving drums with holes in them. From this demonstration it is clear that our visual system has remarkable powers to recover shapes of unknown structures simply from the way their shadows change. That is, from motion.

---

19 For Ullman's work see: "The Interpretation of structure from motion. Proc.R.Soc.Lond." B 203, 1979a, pp405-426; and his book TheInterpretation of Visual Motion. (Cambridge,Mass.:MIT Press 1979b).



Ullman's rotating cylinders demonsrtsation. Dots painted on the two cylinders are projected orthographically onto a screen as indicated by the arrows, giving a sequence of frames like those illustrated below. Each single frame has the appearance of a set of random dots, yet when seen as a movie, the rotating cylinders are clearly visible.



The structure-from-motion problem. This set of frames contains three-dimensional information (see above). How are we to recover it?

Figure 2.2

Marr and Ullman found this result important because the visual processing task is so difficult that any lead as to how new information about the world can be systematically and reliably extracted from visual data would help them understand how the complicated transitions are made from apparently retrograde retinal images to sharp representations of a three-dimensional world.

According to their general theory, visual processing has three cardinal phases. First, the intensity arrays on the retinal cortex are transformed into a two-dimensional image of pictural entities such as points, lines and regions. These primitives are built up in stages in a constructive way, by representing intensity changes, adding representations of the local geometrical structure of their arrangement, and then by operating on these things with rules. Second, this two dimensional image is then transformed, again by a set of constructive processes, into a two-and-a half dimensional sketch of scenes consisting of surfaces, three-dimensional edges, and so on. Finally, a three-dimensional representation is constructed which locates fully-fledged bodies in an object centred co-ordinate system. In each phase, a transition between a representation and a successor representation is achieved by performing certain computations which recover increasingly more of the information about the visual properties of the world.

Now, plausible as the approach is, it runs into difficulties in explaining how these transitions are possible. Construed as purely constructive processes, computations ought to add nothing to the information originally presented to the retina. All the information needed for sight ought to be present at the retina. The goal of

computation is to capture, enhance, resolve and abstract from the complex patterns of neural activation both systematically and sequentially until a new system of patterns is generated. These final patterns, the three-dimensional representations, are in a form that is perspicuous for humans; they convey the right sort of information to allow us to plan and execute our actions so as to achieve our goals. But in principle they ought to contain 'nothing more' than what was presented to the eye. The trouble is that this job proves impossible. Constructive processes can capture a surprising amount of information from low level representations but not enough to uniquely specify the representations we actually end up with. That is, in treating the transitions from two-dimensional images to two-and-a-half and finally to three-dimensions as information processing problems we find that the solution at each stage is underdetermined. The relation between two-and-a-half and three-dimensional representations, for instance, is many-many. There are an infinite number of three-dimensional configurations that could give rise to the same two-and-a-half dimensional images. So we cannot single out one three-dimensional configuration as the one world state causing a two-and-a-half dimensional representation unless we know something more about the world. Since we all do find the solution we must know something about the world that is not derivable from the stimulus conditions alone. This extra something may be as simple as knowledge of past stimulus conditions, but it may also include knowledge of general principles such as:

1) a given point on a physical surface has a unique position in space at any one time[20]

2) matter is cohesive, it is separated into objects and the surfaces of objects are generally smooth[21]

3) the things in the world that give rise to intensity changes in an image are

   i) illumination changes, which include shadows, visible light sources and illumination gradients;

   ii) changes in the surface reflectance;

   iii) changes in the orientation of distance from the viewer of the visible surfaces.[22]

4) structures in the visual world are rigid or at least nearly so.[23]

As Marr put it

> we must bring additional information to bear on the problem that constrains the solution one finds. This additional information must at the same time be powerful and true but rather unspecific. Powerful because it forces a solution that is usually unique; true because not only does one perceive only one solution, but that solution is also the correct one physically; and unspecific because the system works in unfamiliar situations, without specific a priori knowledge of the shapes to be viewed.[24]

Once we have identified this extra information we can incorporate it into the design of the system so that the system will be able to recover more information from the sensory given than was actually there.

---

20 Vision p.113.
21 ibid. p.113.
22 ibid. p.69.
23 ibid. p.209.
24 ibid. p.206-207.

Returning to Ullman's example now, let us see how he shows that the rigidity constraint -- the assumption that most of the structures in the visual world are rigid or at least nearly so -- plays an essential role in recovering structure from motion. This exercise is worth going through because it is hard to see how one could establish that it is necessary that a given principle or fact be known. Prima facie there are many principles, not equivalent to the ones Marr or anyone else may state, that will enhance computational power appropriately. For even if it can be proven that some extra information is required to perform the computation it doesn't follow that it must be the information Marr claims. Admittedly, I cannot think of even one other principle that might replace any of (1) to (4) above. But my inability to think of alternative principles is no proof that Marr's choices are necessary.

To summarize, Marr's argument is that there exists some computational function c_1 which is such that when it is applied to the conjunction of Rep_1 and extra information I_1, it uniquely determines a new representation Rep_2. That is,

$$Rep_2 = c_1(Rep_1, I_1)$$

If I_1 is not available, however, $c_1(Rep_1)$ cannot uniquely determine Rep_2. Hence (c,R,I) must be sufficient for Rep_2.

It ought to be evident, however, that to prove necessity we must show that there is no other combination of computation and information such that $Rep_2 = c^*(Rep_1, I^*)$. That is, we must show that there is only one way of solving the structure from motion problem.

Marr's actual argument can be reconstructed as follows:

A) There is no function c powerful enough to map Rep_1 (the two--dimensional image) onto a unique Rep_2 (the two-and-a-half dimensional image) for all values of Rep_2.

$$c(Rep_1) \neq Rep_2$$

B) A unique Rep_2 does, however, supplant Rep_1. Hence some c must exist.

C) If A) and B) then Rep_1 cannot be a proper description of the argument of c. There must exist some extra information which when added as input to c allows the computation to uniquely specify Rep_2.

$$c(Rep_1,X) = Rep_2$$

D) This extra information amounts to the knowledge that, for example, objects are rigid. X = I. For it is provable that $c(Rep_1,I)$ is sufficient for all Rep_2.

E) I is not only sufficient, however, it is empirically necessary, because it is the only knowledge which it is reasonable to assume the system has. We can explain why the system has knowledge of I. It is not accidental that X = I; it is causally necessary that X = I.

Let me go through this with Marr's (following Ullman's) actual example. It will be recalled that Ullman's chosen Rep_1 was a snow flurry of dots projected from two revolving cylinders. From experimentation we know that all sighted people can infer from Rep_1 the presence of two cylinders, Rep_2. Moreover, they infer the presence of the same two cylinders. Evidently Rep_1 has three-dimensional information that we can uniquely recover. Therefore B) is empirically established. If no assumptions about the world are made, we cannot decide which dots to group together as parts of a

projection. It is impossible to recognize wholes (patterns) if we have no principle for grouping parts together. The question is: Are these grouping principles ones which depend on assumptions about the world? Or can they be defined over local portions of the representation -- for example, according to some principle whereby dots are grouped together if they appear to share a common fate. Unfortunately, the dots at the edge of each cylinder slow down relative to others and appear not to share a common fate with others at all. Nor are there other low level properties that can be used. It is hard to see how $c(Rep_1) = Rep_2$ could be solvable given the constraints on information extraction. Hence A) seems true. But if A) and B) are true C) follows: there must be some extra assumptions made about grouping. Perhaps it is rigidity. As it works out the assumption of rigidity is provably sufficient for the grouping problem, because given the axioms of Euclid it is possible to demonstrate that for any rigid object, four non-planar objects glimpsed with both eyes from three different perspectives provides enough information to allow ideal geometers to infer the three-dimensional shape of the object. If we assume the projections to be projections of rigid objects we can organize the dots into natural groups which provide the two-dimensional information necessary to recover the three-dimensional shape. The assumption of Rigidity is sufficient to allow Rep_2 to be derived. Yet is it arbitrary? Do we think the system makes use of the rigidity asumption just because it works? Clearly not. It can be defended on the grounds that objects in visual space do tend to be relatively rigid (at least rigid for the time needed for three viewings). This is a feature of the world that an adaptive system could exploit. Hence we may expect that nature's

choice of a visual system will be one that relies on rigidity to solve otherwise insoluable problems. Indeed it is hard to see how nature could choose otherwise. Rigidity is empirically necessary and logically sufficient. Q.E.D.

Now if this line of reasoning is correct we have a model of how we can decide whether a given body of knowledge is necessary for a given ability. It has three steps. Step 1: We test to see if some knowledge is required for performing the task. We show that no set of information extracting procedures can deliver the representation we assume we require. Step 2: we try to prove that the knowledge we postulate is actually sufficient for the job: information procedures applied to initial information plus new information are sufficient to yield the requisite representation. Step 3: We show that attributing that knowledge to a system is <u>empirically</u> credible. This last step is critical. It shows that necessary analyses are not purely logical analyses of what any system, whatever its history, constitution, or substrate, requires for its competence. Necessary analyses are empirical analyses.

Let me explain what I mean by 'necessary analysis' being empirical with the help of an example. We know, for instance, that according to general physical principles, for a system to be able to function it <u>must</u> have <u>some</u> energy source. Energy is empirically necessary for performance. No set of energy maintaining procedures can ever ensure that in a closed system any future state required for system performance can be generated from past states. Additional energy is required. This, I will assume, is so abstract as to border on the tautological. How much new energy a system requires depends on the

energy content of future states. To discover how much these states require also involves empirical study -- though to prove that a certain quantity of energy input will be sufficient to get the system from state 1 to state 2, is a simple mathematical analysis. Empirical necessity has thus entered our thinking twice. Once in assuming a system must have an energy source, a second time when assuming that the energy source must be a certain quantity. It enters a third time when we try to discover what kind of energy is required. To decide on such details involves studying the system's constitution, again an empirical inquiry. Does it use chemical fuel or electricity? Particular systems need particular energy sources. Cars and aeroplanes need combustible material, television sets and radios need electricity. The more specific we are about the constitution of a system the more specific we can be about the form of energy it needs. In each cases, however, we feel justified in claiming empirical necessity.

As with energy so with knowledge. Step 1: Analysis of system behaviour may reveal that certain kinds of knowledge are essential for system functioning. In the case of semantic ability truth condition theorists argue that knowledge of the reference of terms is required. This is information which the system cannot extract from its current input, which is phonemic, but which it must have if it is to move to later states of semantic processing.

Step 2. Packages of extra information can be tested for sufficiency. Again returning to semantic processing, is a given set of truth axioms sufficient for linguistic understanding? Are there sentences which we clearly do understand (e.g. metaphorical sentences) which cannot be assigned an appropriate truth condition on the basis of

the given axioms?

Step 3. Can we prove that a given packet of information is empirically necessary: that it is the only extra information it is reasonable to suppose the system must have? In general, empirical investigation is necessary to discover what information (knowledge) a system may have learned, or inherited, or discovered. There must be an explanation of why and how a system has this extra information. In principle, there is no reason why a single information packet may not be selected as the empirically necessary packet. Perhaps systems like S just run on P-type information packets, the way cars run on petrol. Having this information is empirically necessary.

Marr, more than any other cognitive scientist, recognised the potential value of these high level questions. Rather than search for particular process explanations of phenomena, he set out to discover the principles underpinning the processes. In vision, processing occurs in an orderly manner because there are strict precedence relations inherent in the task. The causes of these precedence relations are to be found in general principles of the visual problem. A system which knows these principles is at least in a position to compute a description of the world from a retinal image of it. To be sure, an actual description cannot be computed unless the system has certain computational resources. But the kernal competence rests in its knowledge.

Knowledge of Sufficient Conditions

Marr's work provides a plausible account of what it is to give an analysis of knowledge that is necessary for a given competence. Yet in

fact most of the efforts of Marr and his associates have been directed at proving the sufficiency of bodies of knowledge. In Ullman's structure from motion study, the evidence for sufficiency was mathematical. For most tasks, however, proving the sufficiency of knowledge is more difficult.

Returning to linguistic studies, when a philosopher or cognitive scientist sets out to specify an ability in knowledge level terms, he will, if cautious, state his theory in terms of the propositions that should be known by a rational agent. Thus in presenting his theory he will say that he has discovered a body of propositions, K, which, if known by a rational agent, is such that it will endow him with the ability to perform T. Of course, few theorists actually bother to make explicit the clause 'if known by a rational agent'. Nonetheless it must be assumed. Propositions by themselves are inert. Only when implemented in a system that correctly interprets them do they endow a system with ability. Hence knowledge drives performance only for systems with rational powers. This idea is sometimes minimized. One frequently hears in Artificial Intelligence: "Knowledge is power. The strength of the system is to be found in its knowledge base". But a knowledge base is powerless without an interpretor. The correct account of the architecture of a knowledge-rich system then always distinguishes the knowledge base (the rules base), from the control (interpretor, logic engine).

It is important to make explicit the role of a logic engine, for it is possible to measure the adequacy of a given knowledge base relative to a certain task only if we know the kind of logic or rationality the system has. For instance, following Davidson, we might argue that to

be able to speak a language one must know the satisfaction conditions for all names and predicates in a language, the truth conditions for logical connectives, and have truth conditional analyses for all other operators. Thus, in a Davidsonian theory of meaning we find semantic knowledge described as follows.

'John' refers to John.

'red' refers to $[a_1,...a_n]$.

'p'$^\frown$'and'$^\frown$'q' is true iff p is true and q is true.

'$\Box$'$^\frown$'p' is true iff p is true in all possible worlds.

Are these axioms sufficient for semantic competence as they stand? Obviously not. Only if we can perform whatever logical operations are necessary to apply these axioms can we infer the truth conditions for any sentence the theory was intended to cover. Hence, if I am told "C'est seulement en Marseilles qu'on trouve la vrai Bouillabaisse" I understand the meaning of that sentence, according to Davidson's theory, only if I can apply my axioms systematically and arrive at an interpretation of the truth conditions of that sentence. To succeed in this effort I need deductive ability. I must be able to construct the necessary truth derivations.

The adequacy of a given truth conditional semantic theory is measured against the requirement that for any sentence, s, in the language in question, L, the knowledge base (i.e. axiom list) when coupled with a deductive engine is capable of generating a theorem of the form "s is true iff p", where 's' names the sentence in question and 'p' is a statement of its truth conditions. Thus for every sentence of L, I must be able to use my knowledge base to generate an interpretation

of the form "C'est seulement en Marseilles..." is true iff it is only in Marseilles that real Bouillabaisse is to be found. If there is even one sentence in L to which I cannot assign a meaning in this fashion, then my knowledge base is in some way inadequate. It is missing axioms, and so fails to provide me with sufficient knowledge to know the meaning of all sentences in L.

How can I verify that a given knowledge-base is sufficient? Where there is an independent means of identifying all the meaningful sentences in a language, it is possible to prove the adequacy of a given knowledge-base in the standard metamathematical way. For instance, if all syntactically well-formed sentences are recursively specifiable and every well-formed sentence is meaningful then all we need show is that the semantic axioms allow us to deduce the truth conditions for each grammatical sentence. But what if the ability we are interested in, is not as well-defined as this and there is no recursive specification of all and only those acts which qualify as manifestations of the ability? In such cases will it not be hard, perhaps impossible, to prove the sufficiency of a knowledge-base?

This is a problem of some import. Even semantic competence is not free from this difficulty. I do not propose to question whether it is in fact possible to specify all and only meaningful sentences on syntactic grounds alone. It is easy to construct examples in which syntax cannot be analyzed without some concurrent semantic analyses. So the two are unlikely to be sufficiently autonomous to allow syntax to be used as an independent means of identifying all the meaningful sentences in a language. Rather I wish to point out the vagueness inherent in complex abilities, and the artificial closure one must make

on what counts as relevant knowledge.

Let us consider vagueness first. In low level vision, vagueness about the specification of processes is not a central issue. Marr introduced a rigorously mathematical framework in which computational problems could be raised. Thus Ullman's geometrical proof served to establish the sufficiency of the rigidity assumption because he knew precisely what the rigidity assumption was supposed to allow the system to do. Outside of low level vision however, the situation is far less clear. What is the final output of vision, for instance? Marr said we must be able to recognize objects and locate them in space. But notoriously the process of recognition involves conceptualization. Not more than fifty years ago one of the central questions in philosophy concerned the role of interpretation in vision. Philosophers divided up into sides: those who took visual processes to contain a core 'sensual' element, logically separate from all interpretative overlays, and those who took the gestalt element in vision as primitive, impossible to dissociate from the basic 'contents' of visual experience. Marr has shown that computationally there is a core ingredient in vision. But if vision is to merge with conceptualization as the Gestaltists wish, then conceptualization may be part of the visual process itself. Consequently the formulation of what the visual system does will become as vague as specifications of what conceptualization adds. Should we grant that it is through vision that we can identify a table from any angle; that is, should one maintain that it is vision which enables us to see virtually any kind of table as a table? Or a mountain top as cold? Moreover, how perfect must our sight be for us to have perfect visual competence? Should we be able

to see in the dark? While drunk? While thinking about philosophical problems? We are told: "sight enables recognition under normal conditions". But what are they? Must we be able to exhaustively specify what makes conditions normal? Until we know what the capacity of our visual system is we cannot prove the sufficiency of a knowledge base. Again consider language. The semantic theory for a language L includes axioms for all the terms in use. Does anyone actually know all the terms? Having a narrow vocabulary limits the range of one's speech but there are plenty of fluent speakers with small vocabularies. How large a vocabulary is necessary for linguistic mastery? If knowledge of the whole lexicon is not necessary for being a fluent and competent speaker of L then knowing the complete semantic theory of L cannot be necessary for linguistic mastery either. How much of the theory must I know? How much is sufficient for mastery? There is no precise answer. Yet without an exact formulation of linguistic mastery, no knowledge-base can be verified as sufficient. We have no means of deciding how much knowledge is enough to guarantee competence. The standard solution is to just stipulate a certain level of lexical knowledge that is sufficient for linguistic mastery. But this introduces the second obstacle to sufficiency proofs; the problem of circumscription.

To characterize this problem I shall need a new example. I shall use McCarthy's discussion of the puzzle of the missionaries and the cannibals.[25]

> Three missionaries and three cannibals come to a river. A rowboat that seats two is available. If the cannibals ever outnumber the missionaries on either bank of the river, the missionaries will be

---

25 John McCarthy, "Circumscription -- A Form of Non-Monotonic Reasoning", reprinted in Webber and Nilsson (Eds.) op.cit. p.467.

> eaten. How shall they cross the river?
>
> Obviously the puzzler is expected to devise a strategy of rowing the boat back and forth that gets them all across and avoids disaster....
>
> Imagine giving someone the problem, and after he puzzles for awhile, he suggests going upstream half a mile and crossing on a bridge. "What bridge?" you say. "No bridge is mentioned in the statement of the problem." And this dunce replies, "Well, they don't say there isn't a bridge." You look at the English and even at the translation of the English into first order logic, and you must admit that "they don't say" there is no bridge. So you modify the problem to exclude bridges and pose it again, and the dunce proposes a helicopter, and after you exclude that, he proposes a winged horse or that the others hang onto the outside of the boat while two row.
>
> You now see that while a dunce, he is an inventive dunce. Despairing of getting him to accept the problem in the proper puzzler's spirit, you tell him the solution. To your further annoyance, he attacks your solution on the grounds that the boat might have a leak or lack oars. After you rectify that omission from the statement of the problem, he suggests that a sea monster may swim up the river and may swallow the boat. Again you are frustrated, and you look for a mode of reasoning that will settle his hash once and for all.

McCarthy was illustrating a deep and unsolved problem in Artificial Intelligence: the frame problem. The frame problem arises in problems set in real-world contexts where one needs to know which environmental features are relevant to solving the problem. People carry vast amounts of information about what will be changed by actions and what will not; about objects and features of the world that can be relied upon to behave in certain ways and so harnessed to ends, and about what can be expected to arise in certain situations, and what can be ignored -- in short, what is normal. In order to solve a problem in real life it is tremendously difficult to state what information may be required. Indeed the point of McCarthy's example was that there is no principled way of circumscribing the class of facts that could be relevant to a

solution. Normally, relative to a particular formulation of a problem there is a class of canonical paths that will lead to the solution. These are the paths consistent with the spirit of the problem. Accordingly, if the agent has the right knowledge and he has the requisite logical machinery, sooner or later he will succeed in finding and following one of these paths. And in formal deduction this is exactly what we find. Given a set of search heuristics -- as a knowledge-base, and given a set of deductive rules and a means of applying those rules -- as logic engine, we can expect that sooner or later the system will find the solution to most any theorem. There are a class of canonical paths to the theorem, called proofs, which hold good for all time. Adding premises never diminishes what can be proved from the premises: additional axioms never invalidate the proofs of earlier theorems. The paths are stable and reliable. Tacking on new axioms just increases the number of proof paths. Hence relevance is circumscribed to whatever axioms or theorems might appear in a given class of proof paths, and sufficiency is provable by proving the adequacy of (in this case) the heuristics (or the principles on which they are based).

In non-deductive contexts, however, particularly in contexts where practical reasoning is involved, complications arise. Who is to say that in the real world goal-oriented paths are reliable? In the abstract they may seem sufficiently 'on course' to guarantee that anyone following them will reach the goal state. Yet in real life "things come up". The unanticipated occurs and even the best laid plans fall through. How much knowledge must one agent have about possible pitfalls and interferences to be able to extricate himself from any

situation and bring himself back on course? Unless circumscribed, this problem soon gets out of hand. But this is just the sort of problem real life poses.

It is characteristic of living systems that they can adapt to local perturbations in their environment. They cope. Slight inaccuracies in the data presented to the eye do not throw the eye so far off course that it fails to complete its computation. Allowance is made for noisy data. The system is able to deliver an account of the object that, though rough at first, soon becomes increasingly accurate as more views and hence more information are presented. Similarly, in plan-making,it is a hallmark of a system's rationality that it can find new routes to its goals, it can innovate. Unless people could be counted on to do whatever is necessary to overcome unsuspected obstacles in their path, they would not be sufficiently trustworthy to be included as participants in bigger projects. People adapt to the unanticipated and can be relied on to do so.

This feature of living creatures is essential to their abiltiy to survive. A system which lacked flexibility in its reactions to situations is not likely to last long in a world of uncertainty and change. Since all real world situations we know about involve a measure of uncertainty, competent systems must be endowed with sufficient knowledge to cope with at least some environmental perturbations. The question is: How much? How much knowledge is enough knowledge?

Returning then, to the cannibal/missionary problem, how much knowledge must a system have to truly be said to have the ability to solve the problem? Is there a unique answer? Everything hinges on whether it must be able to solve the problem under any perturbation.

Clearly there must be some restriction on variations. But where is the line to be drawn? Can we say that some environmental changes change the problem, and hence change the ability required to solve it?

Suppose, for instance, as soon as I think I've found a solution to the cannibal problem my questioner adds "I forgot to tell you: the rowboat has no oars.". My solution is invalidated now. It is incomplete. How can people get from one side of the river to the other in a rowboat without oars? There is a new problem now, one requiring new knowledge. I need to acquire the ability to solve another problem.

Yet isn't there a sense in which the problem is the same problem, only qualified -- I don't need a new ability at all? The situation can be likened to a runner who claims to be able to run four minute miles. He is taken to Mexico and can't break four minutes ten seconds. Does he have the four minute ability? He argues he does. The track in Mexico is at high altitude. He didn't say he could run a four minute mile so high up, he meant a four minutue mile at sea level. We take him now to the Ivory Coast and he tries again. Once more he fails, this time saying he meant he can run four minute miles in temperate conditions. The situation is repeated. For each trial which apparently disproves his ability he qualifies his claim. To stave boredom and forestall his tormentors, in the end, he will just subsume all further <u>interferences</u> as violations of normal conditions. His real ability, he informs us, is to run a four minute mile under <u>normal conditions.</u>

We have come to the heart of the issue. Sufficiency analyses tell us what information a rational system will find sufficient to enable it to perform T under normal conditions. But they leave the meaning of

'normal conditions' implicit. As I shall argue in Chapters Four and Five, difficulties in specifying normal conditions threaten the prospects for effective knowledge analyses of our higher cognitive faculties. I think there is certainly a qualitative difference between the vagueness problem characteristic of linguistic competence, visual competence and other 'peripheral' faculties, and the normal conditions problem characteristic of reasoning tasks. In the end, this has to do with the kinds of interferences we feel obliged to acknowledge as relevant. In the case of vision, we can identify a low level component which must be capable of computing three dimensional images from retinal data of a certain sort. The system is nicely closed to interferences from external (and internal) sources because although we define the goal of visual processes in terms of a correspondence between the spatial and formal characteristics represented internally and those found in the external object, we define the computational task over proximal stimuli. The task environment of low level vision is remarkably spartan. The moves in the visual task space are not defined over operations in physical space, where all sorts of interferences might crop up. They are defined in a mathematical space given by the constructive operations that can be performed on the intensity arrays encoded at the retina. Similarly, the moves open to a competent speaker are not transformations, or mirror images of transformations, of physical state into physical state, as was the case in our car assembly problem discussed earlier, or as is the case in problems to do with action. They are transformations of syntactic or semantic structures into syntactic or semantic structures. The constraints are restrictions on allowable transformations. And we can

consider these without much concern for the 'normal conditions' of the task space.

Such abstraction from 'real' environments may undercut the intuitive urge to call high level characterizations of the linguistic and visual faculties, characterizations of knowledge. For classically, knowledge is something that comes in through the senses and participates in guiding worldly actions. But, as I shall argue in the next chapter, there are other grounds for calling such high level charaterizations knowledge. All that need be appreciated is that the knowledge that constrains processes in our linguistic and visual faculties is not the same sort of state as that involved in processes of higher reasoning: one is tacit, the other is occurrent. But that is to rush the story. In the next chapter, I shall defend the claim that knowledge states are real.

Chapter Three

IS KNOWLEDGE REAL?

Philosophers divide into two groups: those who see the knowledge level as a pragmatic necessity, unavoidable in making predictions about the behaviour of complex systems but incapable of delivering theories that are literally true; and those who see the knowledge level as a robust level of scientific analysis no different in methodological status than, say, Mendelian genetics, functional ethology, or economics, and capable of generating theories that are as likely to be literally true as theories in these other disciplines. The position I am defending is that knowledge states are real states of a system, states which, in some sense, are causally effective and which are relatively permanent features of the system. Theories about knowledge may be literally true.

In this chapter I shall reply to some skeptical attacks on the idea that knowledge states might be real, intrinsic states of a system. It has seemed preposterous to some that anyone could claim that knowledge states, which necessarily are ascribed to a system only relative to a certain task environment, might nonetheless be intrinsic features of knowledge-rich systems. In their view, knowledge is essentially a relational state. Hence it need have no categorical basis. If some object A is **beside** object B, there is nothing intrinsic or categorical in A in virtue of which it is beside B. Being beside is a relation that is purely extrinsic; it applies to any pair of entities in virtue of their spatial relation. Knowledge, it is suggested, is

like that. In fact, in the case of knowledge, its relational character is even more confused, for different observers allegedly may ascribe different knowledge states to the same system. Knowledge, we are told, is observer relative as well as environment relative. How then could we expect our knowledge attributions to set constraints on design? Constraints would vary from observer to observer, environment to environment.

The objection has substance. It is well known that given enough freedom in interpreting behaviour, any system can be interpreted as knowing virtually any body of propositions. Just as the lines in an unused notebook can be interpreted as the words of King Lear -- provided that we have been given a bizarre enough interpretation manual -- so an inert block of gold can be interpreted as knowing Shakespeare's sonnets, or Peano's axioms, for we can, if sufficiently creative, read its responses to light changes, heat variations and the bare passage of time, as displaying sophisticated linguistic or arithmetic behaviour. Accordingly, no interpretation will be compelling unless we have evidence that the interpretation manual we adopt is non-arbitrary. In the case of intentional action and linguistic behaviour in particular, it is questionable -- or at least so some have argued -- whether such evidence can be found, whether there is a fact of the matter about which interpretation is correct. This then is the first challenge to knowledge level research: Why think knowledge attributions are ever true since we cannot tell arbitrary from non-arbitrary interpretations?

Cognitive Ethology

Against this I have been arguing that the possibility of arbitrary interpretations, though real enough, can be avoided if we have an unclouded idea of a natural task environment. It is not an accident that we ascribe the visual faculty knowledge of certain visual invariants, knowledge of empirically true generalizations such as that medium-sized objects tend to be rigid. Our visual system was specifically designed to take advantage of truths of the visual world. The same may be said for our reasoning faculty — if it makes sense to see it as a single monolithic faculty. We are able to draw rational inferences because we have been designed to see connections between propositions, to see that C follows from A → C, A; or to see that certain actions can serve as means to ends. The evidence that our ability to draw such inferences is not an arbitrary interpretation imposed without foundation but is rather a designed feature of our minds is less easy to discover than evidence about the non-arbitrariness of our interpretation of visual processes. And yet even here there is a case to be made for treating certain behaviour and certain inner processes as genuinely about states of the world, states of a planning space, a task environment.

Those who see the reference relation as a natural kind relation, for instance, suppose that the reference of certain symbolic structures, is not a mere accident; whether the symbols are external linguistic ones or mental or neural ones, it would be wrong to think they have no more grounding in fundamental facts of the world than the length of a meter, or the date of the New Christian Era. Reference, representation and aboutness are all relations which arise through complex causal interactions, both

perceptually and behaviourally, with the objects or states referred to. Work has just begun on such studies.[1] It is too soon to say whether the hypothesis of reference as a natural kind will be borne out. Yet there is hope. For if ethologists can talk of the functions of behaviour, of the specific states or processes in the environment which certain behavioural acts were designed to bring about (or prevent occurring), and if we grant that these ascriptions can be true, then why not grant that intentional ascriptions can be true of selected behaviour and cognitive processes? Both are ascribed relative to an environment. Both require interpretation. But both presuppose networks of causal relations that nomologically link the processes or states in question with particular environments. Thus if we can feel confident that white headed gulls move broken egg shells from their nests because broken shells are easily spotted by the sort of eagles in their environment, why can we not also feel confident that certain (linguistic) behaviour means 'all men are mortal', or that certain internal changes in the cognitive system occur in order to ensure that 2 + 2 is correctly summed? Is attribution of functional role[2] different from attribution of representational role?

---

1 See especially Gareth Evans, Varieties of Reference, (Oxford: Oxford University Press, 1982); and Hartley Field, "Tarski's Theory of Truth" reprinted in Mark Platts (Ed.), Reference, Truth and Reality, (London: Routledge and Kegan Paul, 1980).

2 It must be obvious that the functional role ethologists discuss is not the type of function which psychologists normally discuss. Psychologists are interested in input-output function. To specify input and output it may be necessary to acknowledge environmental characteristics, for instance, to identify an action as throwing a ball we must acknowledge that there are balls in the environment. But much of psychology is concerned with behaviour that can be

Throughout this thesis I am assuming that reference is a natural causal relation. There is a fact of the matter, then, whether a given structure or process refers to some individual or state, or not. We may at first be uncertain about which individual a symbol refers to, but this uncertainty is not inevitable, hard historical research can expose the causal lines leading from a word or mental process to its referent. Consequently, the first phase of cognitive science -- the phase of knowledge attribution and analysis -- is not unlike the first phase of ethology. In both sciences our objective is to find the correct non-structural or relational description of the system before going on to explore possible mechanisms by which the system may satisfy those descriptions. In each case this involves discovering the natural environment of the organism.

In ethology, for instance, the function of a behavioural action is always relative to a niche, to a world with respect to which it was designed to be adaptive. The added phrase "it was designed to be" is not gratuitous. The ethologist knows that behaviour that serves some function F in world W may serve F^* in W^*, just as a bolt may function as a structural member in one machine, a shear pin in another and as a sinker on a fishing line. Since behaviour may, by coincidence, be adaptive relative to a world despite its not being designed to be adaptive in that world, some effort must be made to locate the 'canonical' environment of the system, the original

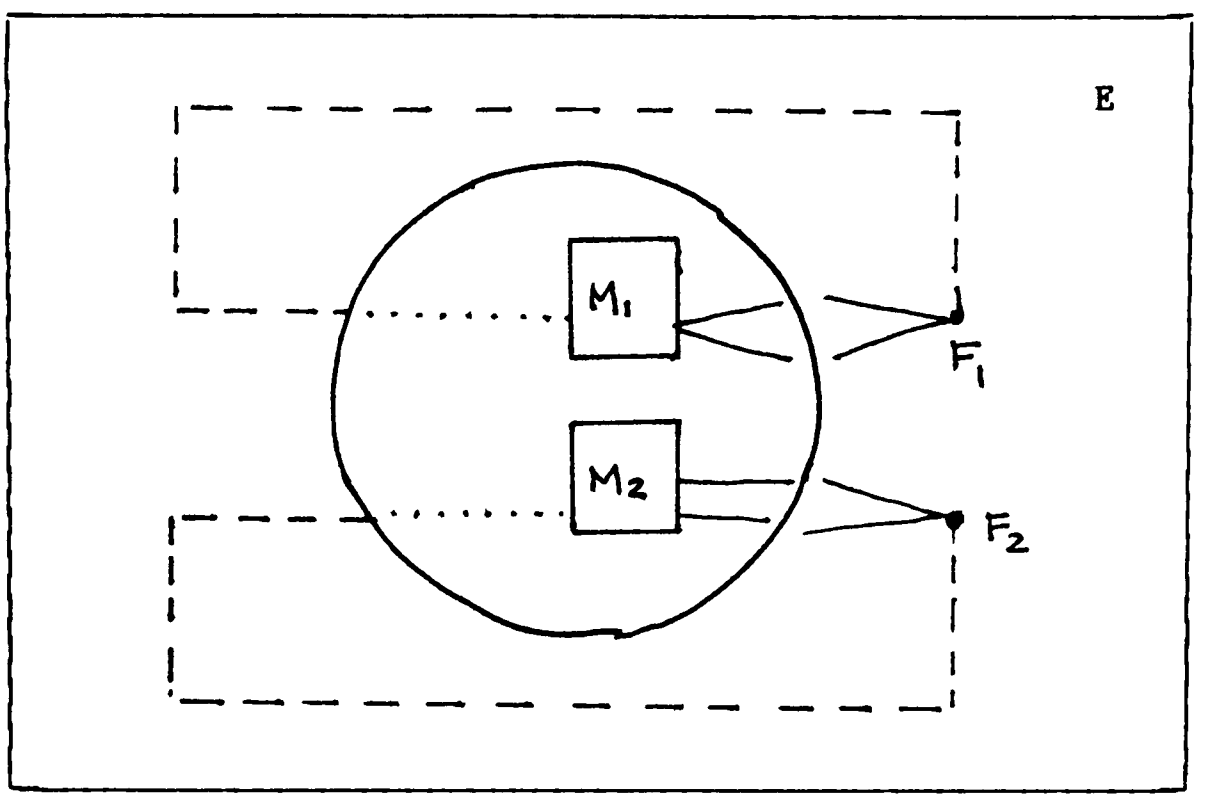
---

structurally identified and so can be discussed without concern for the environment in which the behaviour is performed. In ethology functional role is necessarily tied to objects in the environment. The frog's tongue is designed to catch flies. And there is no way of specifying that function without referring to flies.

environment. We must be careful to identify the original function of the behaviour, for new environments distort the original trade-offs of design which were causally relevant in the process of natural selection, and hence give us a distorted picture of the kinds of influences which would be decisive in designing the mechanism.

For example, in Fig 3.1, the mechanisms postulated to explain observed behaviour depend on the functional role assigned to the behaviour. In environment E, behavioural actions A_1 & A_2 are interpreted as customarily having effect F_1. Accordingly we say the function of A_1 & A_2 is F_1. In E^*, however, A_1 & A_2 have different effects. A_1 characteristically causes $F_2{}^*$. For instance, suppose A_1 & A_2 represent colour changes which the organism undergoes. A_1 will be the change from brown to green; A_2 the change from green to brown. In E both changes may help camouflage the organism against the colour sensitive vision of its chief predator, the eagle. But in E^* where no eagles exist, A_1, may seem to attract grass flies, or other edibles that migrate toward green things. A_2, meanwhile, may frighten smaller predators. Now because of the different functions served by these actions in different environments, and because of the causal structure of the different environments, it may be easiest to collect behaviour into natural packages in rather different ways in each environment. When the system is studied in E and we are asked which behaviour we believe to have a common origin in an underlying mechanism, we choose A_1 and A_2. When the system is studied in E* we may decide to associate A_2 with other behaviours serving an aggressive function. Thus A_1 is caused by mechanism 1, while A_2, A_3 and A_4 are caused by mechanism 2.

Functional Decomposition Depends on Environment



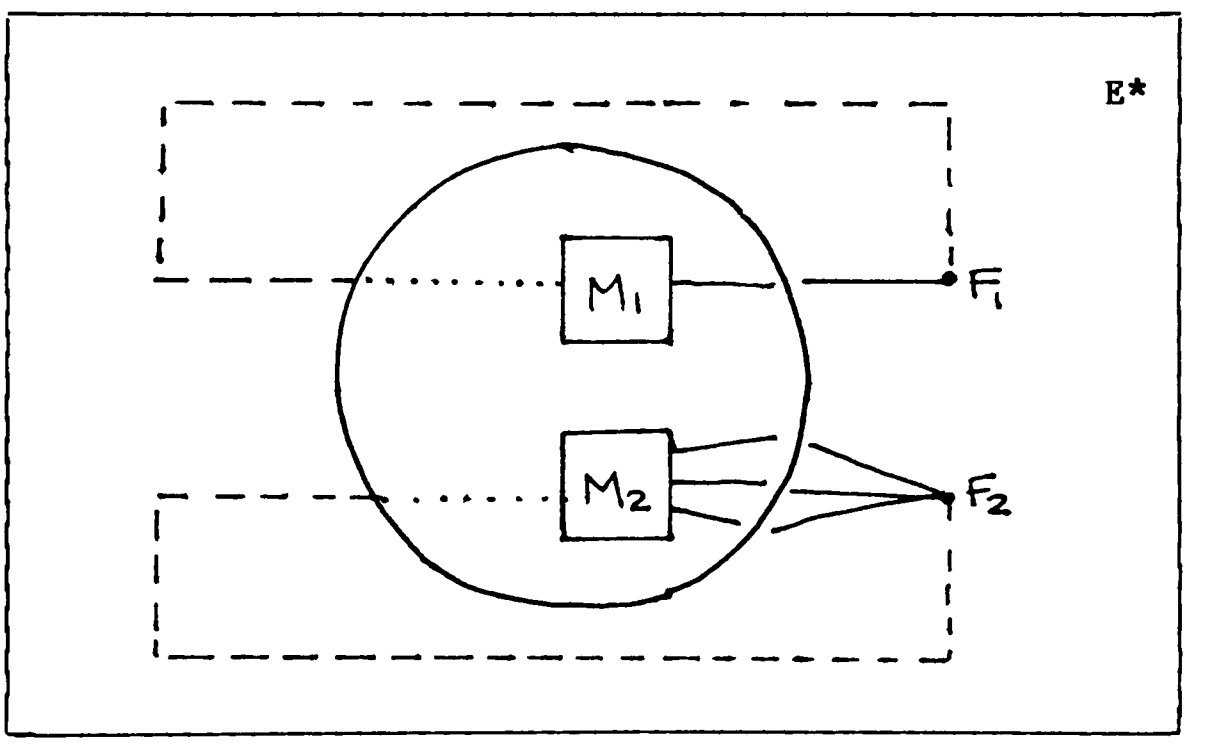


Figure 3.1

In figure 3.1 an organism with a repertoire of four acts is transplanted from its natural environment E to a new environment E*. In E* its behaviour receives a different ethological interpretation. What once was seen as camouflaging activity is recharacterised in E* as luring activity or frightening activity. On the basis of these functional accounts the ethologist speculates on mechanism. The same system, then, may support different hypotheses about internal structure in different environments. F represents the type of state of affairs in the environment that is a type of action's function. Over generations we may see F as being a causally important factor in selecting a design of the system via the mechanism of natural selection.

In knowledge level research, similarly, care must be taken in selecting the natural environment of the organism. Every naturally produced cognitive system is composed of a network of cognitive faculties. Each faculty has been shaped by certain environmental demands tempered by the internal demands of scarce cognitive resources. Within these constraints the faculty has evolved to perform a natural task. If we study in detail the demands imposed by the natural environment of the organism, and analyze what the organism actually does to meet those demands, we may begin to articulate a taxonomy of natural tasks, natural abilities, that are called on again and again by the creature during the course of its daily life.

Where cognitive ethology departs from ethology is in the method of specifying tasks and abilities. The human environment is primarily an artificial environment; man is insulated from the 'brute' world by his own constructions, none of which is more significant than symbols. In modern life a person incapable of perceiving and understanding norms, conventions, linguistic and non-linguistic symbols, would soon perish. Somehow nature has equipped us with the means to recognize, assimilate and react in a reasoned way to symbolic events. Animals, it is widely thought, do not have this ability, or have it to a limited extent. Symbolic events are individuated (primarily) by properties that are non-intrinsic: their meaning is to be found in the entities they designate or represent. We can imagine a rat learning to associate with a particular form of behaviour some normal causal consequences.

Thus, pressing the food bar may have a 'significance' for the rat. But the act's significance as grasped by the rat is contingent on its having noted associations in the past. It is not clear that the rat takes bar pressing as representing food. The meaning of an intentional action, by contrast, is to be found in most cases in the point of the action. Since the point of an action may not be a state in the actual world, or even in the world of physical possibility, but rather in the world the agent believes to be possible, it is not necessary that it normally be causally connected with its point. The capacity to see the point of an action, then, does not depend on having associated the performance of the action with certain of its consequences. It relies, instead, on seeing the end of the action within the world as conceptualized by the agent.

The difference in orientation this produces is captured by a distinction between parametric and strategic and the different forms of rationality to which they give rise.[3] The parametrically rational actor treats his environment as a constant: if it is filled with other actors he behaves as if he is the only one whose behaviour is variable and all the others are parameters for his decision problem. The strategically rational action, however, takes account of the fact that the world is made up of other intentional actors, that he is part of their environment too, and that they know that he knows this, and so forth. Each strategic actor has to take account of the intentions of

---

3 Elster, Jon Ulysses and the Sirens, (Cambridge: Cambridge University Press, 1979), Sec. I.4.

all other actors, including the fact that their intentions are based upon their expectations concerning his own intentions. Agents have expectations about the expectations of others.

The corresponding increase in demand on computation this creates is so large that systems able to recognize and respond to actions may require cognitive architectures that are significantly different from those possessed by intentionless beasts.

In the last chapter I argued that to specify a cognitive component, we should think in terms of the natural task it is to perform. If it serves a general informational function in the cognitive system we shall see it as a faculty which has been designed to perform a definite job in the cognitive economy. To specify that job, its 'natural' task, we must use epistemic terms. We cannot look to behavioural performance alone. A faculty's impact on behaviour is mediated by other faculties and their inner war over the use of scarce cognitive resources. Hence behaviour seldom, if ever, reveals the full potential of a faculty. We must discover what knowledge the faculty was designed to acquire, transfer or modify. Occurrent knowledge is the medium of exchange in the cognitive economy. Hence to define the function of cognitive units we must specify the occurrent knowledge they receive and transmit. Once this epistemic function is grasped we can look for constraints on processing.

Categorical Basis

I think that this entire way of viewing knowledge level research makes sense only if we see knowledge as a genuine property of systems, quite unlike extrinsic relations such as beside or after. It is precisely

because we see knowledge as having a categorical basis in a system that we expect it to constrain design. To be sure we must distinguish between different sorts of knowledge. Occurrent knowledge is not identical with tacit knowledge, virtual knowledge or conscious knowledge. The categorical bases of these different sorts of knowledge may vary: some have a basis that is distributed among lower level processes and highly complex, while the basis of others may be localisable in neural tisssue and 'relatively' simple. Nonetheless any adequate theory of a cognitive faculty will identify the different types of knowledge. See Figure 3.2.

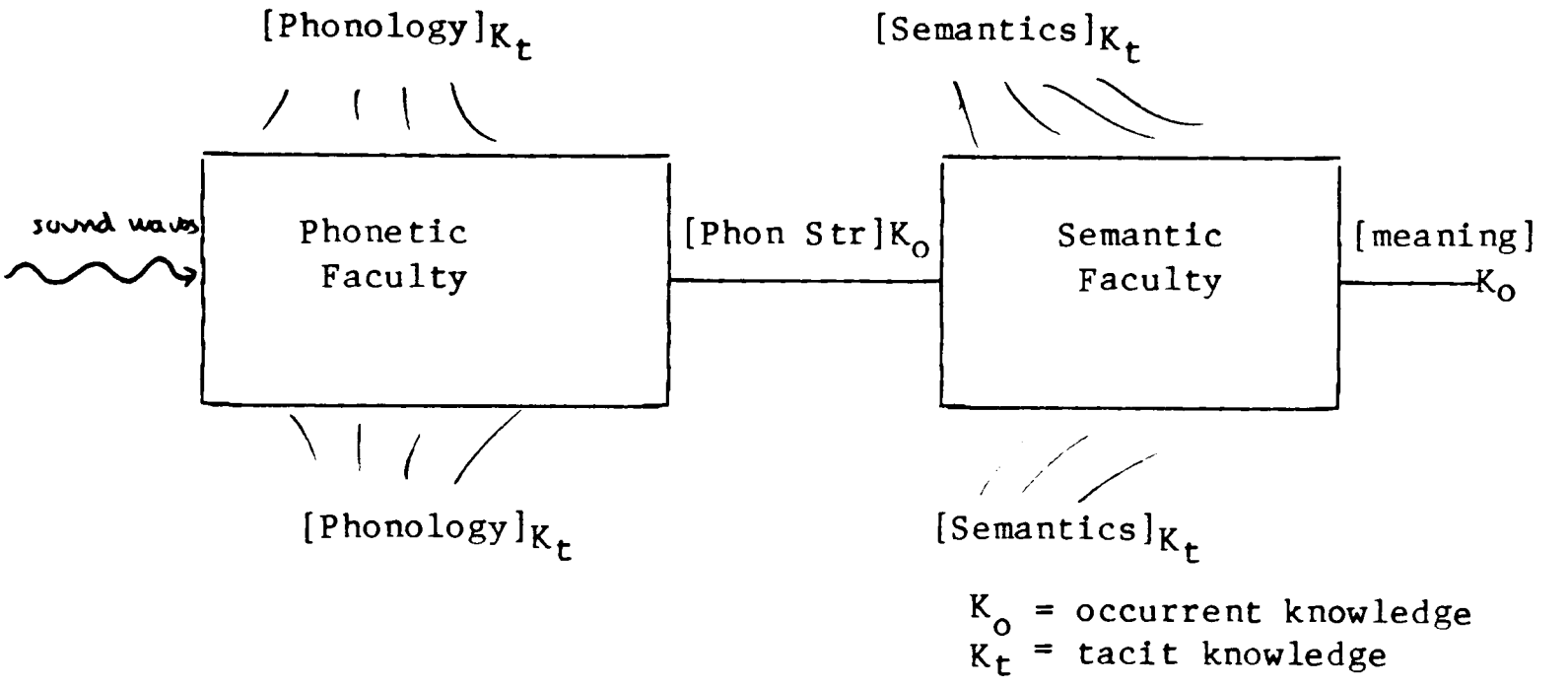


Figure 3.2

Thus if the goal of linguistic communication, for instance, is the transfer of occurrent propositional knowledge, then we know that the channels and receivers of linguistic communication must be structured in a manner suited to propositions. If we treat phonological processing as relatively autonomous from semantic processing then we can characterize the input of the semantic processor as occurrent knowledge of the phonetic structure of sentences heard. Occurrent knowledge is real output; it is a state the system enters that can

influence reasoning. This occurrent phonetic knowledge will then be transformed into occurrent semantic knowledge of the proposition transmitted. Since signals, phonetic or otherwise, produce knowledge of content only if the receiving system has the requisite background knowledge to disambiguate them, the semantic system will have to be designed to have access to whatever background knowledge is necessary for disambiguation. Accordingly, causal pathways to understanding must exploit these background states. How this is done no one can say at present. But these background knowledge states, nonetheless, constitute the standing conditions of understanding. They are cognitive resources that are necessary but not sufficient for understanding and set heavy constraints on the design of any system able to exploit them at speeds required for virtually instantaneous understanding.

This general conception of cognitive faculties might be stated differently. When a creature is well adapted to its task environment it is so constituted that its inner processes generate behaviour that conforms with the constraints implicit in the task. If the system is playing chess then it generates behaviour that obeys the rules of chess. If the system processes visual information, the processes must be structured to produce occurrent knowledge of what is where in the visual environment. Sometimes there are many possible causal pathways to this goal; different systems may have vastly different architectures and support different processes. In other cases, vision perhaps, the the class of different architectures is powerfully constrained and the range of cognitive processes limited. So far this simply recapitulates the uncontentious portion of what has been said before. The

substantive point made earlier was that even in systems of different architecture processing constraints can flow from task analysis if there are clear precedent relations implicit in task structure. Just as in automobile factories certain subtasks must be completed before others can be begun, so it is that cognitive faculties must observe the restrictions imposed on the production line of knowledge manufacturing.

As helpful as production line notions are, it would be misleading to exaggerate what we can infer about lower level processes. Properly construed, such knowledge level characterisations tell us only about the high level states of the processing space. We know that processes must unfold within this space. If knowledge states are real states only mechanisms which support the right system of causal relations will be able to instantiate knowledge. But this doesn't tell us how knowledge states are realised in systemic processes, or how knowledge states constrain those processes. For as the interactions taking place within this space may be between representations, neural structures or whatever, the diversity of possible mechanisms that might successfully operate within a given process space are numerous beyond count. The key tenet of knowledge level research is that although the specific 'shape' of cognitive or neural processes may be different in different organisms, all will share certain global characteristics. Thus although we don't know what neural etc. interactions are occurring in a speech mechanism, we still know that whatever the neural interactions are, they satisfy certain high level constraints.

So far, however, I have said very little about what this constraining relation might be. I argued that it is stronger than supervenience but weaker than the standard nomological ties found in

familiar inter-theoretic bridge laws. Before defending the realist's position it may prove helpful, at this point, if I digress to explain how knowledge might constrain mechanical design. It must be stressed that this speculation on inter-level relations is pure conjecture. Given our present ignorance the best we can do is consider possible analogies of the knowledge/mechanism relation. Invariably analogies are of limited value, particularly fanciful ones. But unless we have an idea, even a rough idea, of how knowledge might constrain mechanism, the following discussion of realism will seem disconnected from scientific research. The computational metaphor provides cognitive scientists with the comfortable illusion that their theories are implementable, that from theorizing at 'high' levels of description they will eventually be able to show that in fact the brain does compute a certain function, or does embody certain beliefs, desires and knowledge states. No one requires localisation of belief etc., but cognitive science is predicated on the faith that one day neuroscientists will show us that the brain is indeed the sort of device that could 'run' the procedural theories of cognitive science; that it does, in fact, link up states by their content, or perform operations that preserve intentional relations. Accordingly, any defence of realism about knowledge that does not allude to the categorical basis of knowledge, neural or otherwise, is necessarily partial. It is for this reason Fodor and Chomsky have always alluded to the 'encoded' nature of knowledge and belief. If we reject the encoded view, but remain realists about knowledge, we must have at least some idea of alternative categorical bases. The three analogies, I propose, do not say much about the nature of these bases, but they

point to additional ways of interpreting the 'constraining' relationship. It is for future science to discover the proper interpretation of knowledge constraint.

How does Knowledge Constrain? Some Analogies

The first analogy of how knowledge might constrain mechanical processes plays on the image of knowledge as a global constraint. Mechanistic descriptions are characterisations of locally interacting structures. They provide us with a detailed account of how and when parts interact As a theory about local interactions, however, mechanistic theory has no need to mention global properties of interactions. Proof of the existence of such properties may help us demonstrate that no design, no combination of local interactions can produce a certain state, or property or structure. But discovery of these global properties tells us little about actual mechanisms. To get a feel for the nature of impossibility proofs, consider the problem of the mutilated checkerboard.

We are asked to imagine an 8 by 8 checkerboard with 2 opposite corners cut out. Assuming we are given dominoes whose dimensions are precisely 2 squares by 1 square, the problem is to decide whether the dominoes can be arranged so as to cover every square on the board. Are there any global constraints here? Is there a certain restriction on possible arrangements? Clearly so. Since the corners of an 8 by 8 checkerboard must be the same colour (for the opposite corners of any even numbered board are always identical) and since a domino is 2 x 1 squares (and so must cover one black and one white square), no set of dominoes can cover all of the squares if there is not an equal number

of black and white squares. Since there are going to be 32 squares of one colour and 30 of another no set of dominoes can cover all squares. The structure of the checkerboard imposes a system of global restrictions on local arrangements. We can prove from the global perspective that a mutilated board cannot be filled with dominoes. We have a proof of the impossibility of any domino laying process ever succeeding.

Perhaps knowledge constrains process like that. Qua structural state it limits the kinds of processes that can unfold in the space it defines. It is like the fabric of space-time. Events take place within it, observing the constraints which are implicit in its structure. The space does not tell us how objects get from A to D. All we know is that anything which does get from A to D must proceed through B and C, though possibly there are further conditions implicit in the topology of the space.

A second possible analogy may be found in the way chemical boundary conditions set constraints on physical interaction. The class of chemical systems, is a proper subset of the class of physical systems. Under specified conditions the physical interactions in a system are so organized that they realize a chemical law. We now know that the reason why collections of physical units behave as stable chemical units is that the degree of freedom of each physical unit is so constrained by the ordering of other physical units in its neighbourhood that all physical units interact in a stable, highly restricted fashion. The laws of chemistry, accordingly, describe recurring patterns that emerge in appropriately bounded physical systems. That is:

Chemical systems = Physical systems when bounded appropriately.

Laws of chemistry = Laws of Physics + Boundary conditions.

Do we have here a model of the constraining relations we are looking for? Might knowledge act as boundary conditions for, say computational systems, differentiating the class of computational systems that are adapted to certain cognitive tasks from those that are not? In that case:

Cognitive systems = knowledge-rich systems = computational systems when bounded appropriately.

Of course, we couldn't expect a comparable law statement to the effect that:

Laws of cognitive systems = Laws[4] at the knowledge level = Laws at the representational level + Boundary conditions.

for there are no process laws at the knowledge level. We have principles and precedence relations, but no process laws. Nonetheless, that should not stop us from trying to discover the nature of the boundaries which possession of knowledge imposes on computational processes.

It is no argument against this view that at present we have no idea of how to draw such boundaries. Chemists had no idea of how to state the boundary conditions of chemistry prior to the development of

---

4 cf. Allen Newell, "Knowledge Level" AAAI Presidential address, CMU publication CS-81-131. "The knowledge level permits predicting and understanding behaviour without having an operational model of the processing actually being done by the agent." p.24.

physical chemistry. Biologists today are in the position of chemists a generation ago. They have, at best, an inkling of the kind of biochemical boundary conditions necessary for stable biological processes in higher mammals, but higher mammals are but one biological type. The space of possible biological systems is virtually uncharted. Moreover, because non-biochemical parts (e.g. synthetic hearts etc.) may serve as functional equivalents in biological systems, the proper boundary conditions on biological systems may not be biochemical. It seems an accident of nature that all the biological systems we know are also made from biochemical parts. It is more likely, then, that the boundary conditions will have to be stated in chemical terms and applied to chemical systems. Thus we may expect that

Biological systems = Chemical systems when bounded appropriately

It will be a matter for chemical biologists to tell us what the boundaries are.

The same, I suggest, holds for cognitive scientists. Dedicated research will reveal the nature of epistemic-functional or epistemic-structural relations. This may sound utterly mysterious at present. Yet consider: a computational system is made up of two logically distinct parts, an unprogrammed 'base' machine that is capable of following any program written in any appropriate language, and a program.[5] See Fig 3.3.

---

5 For an excellent discussion of the logical distinction between the 'base' machine and the program (or structural field) of computational systems see Brian Cantwell Smith, The Computational Metaphor, (unpublished draft, 1982).

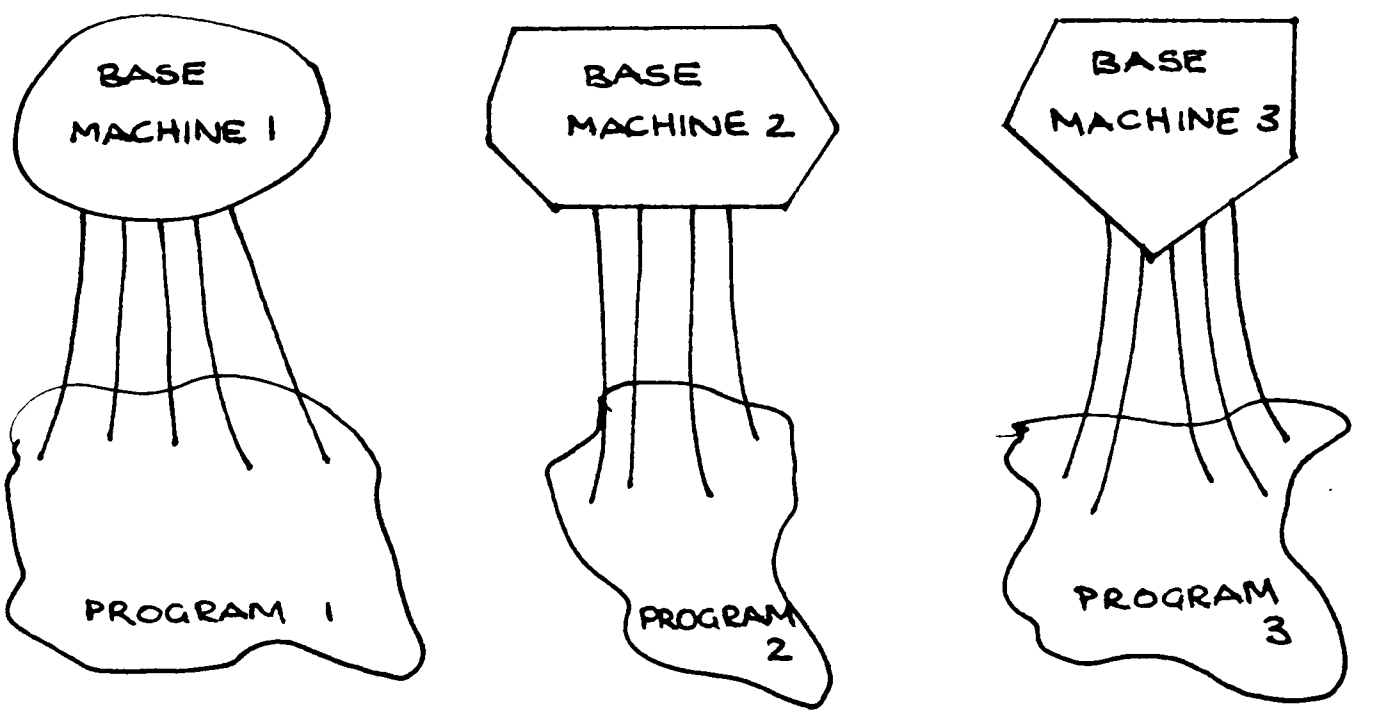


Figure 3.3

Changes in the base machine can be accomodated by changes in the program and vice versa, so that two systems operating with different base machines, and syntactically different programs, may yet be input-output equivalent. Yet are all base machines equally likely to be suited to support cognitive states? Are all equally the stuff from which cognitive systems are made? At present we have no way of deciding this question. But I think I am not alone in assuming that only some are suitably designed to be cognitive systems. The reason, of course, is that only some bear the right sort of causal connections to their natural environment to be truly ascribed knowledge. As we advance in our understanding of the requirements of attributing knowledge, research at the knowledge-level may help us to discover conditions which the base machines must meet. Perhaps they will have to be able to learn, or to perceive. But once we have constraints on the base machinery -- on the fundamental architecture of the mind --

constraints about the sorts of programmes that can be run on those machines soon follow. This is no doubt optimistic in the extreme. But if we cannot hope to discern constraints on the base machinery of cognition for all systems, Martian as well as human, at least we may advance in our understanding of the human base machine. To achieve this understanding cognitive scientists will have to work with neuroscientists. But the same was once true of physicists and chemists: they had to work together to discover the constraints which chemical systems imposed on physical systems.

The third and final analogy of the knowledge/mechanism relation is potentially the most fruitful, but also the one I know least about. In causal embryology research attempts to discover how embryonic growth is controlled. So far, the mechanisms are little understood. Apparently contributing factors are distributed throughout the system and co-ordinated in some way. As J.Z. Young remarked, "there is little detailed information about this distribution or how it is controlled; probably it is by both systemic and local factors".[6] Nonetheless growth unfolds under constraints. Indeed it is the belief of many embryologists that these constraints can be interpreted as higher co-ordinating principles whose operation can be described mathematically (namely, in the formalism of catastrophe theory).[7] It is this possibility which may make the embryological analogy the best analogy of the way cognitive or perhaps neural processes unfold under the influence of knowledge.

---

6 An Introduction to the Study of Man, (London: Oxford University Press, 1974) p.209.

7 For a readable introduction to catastrophe theory see Alexander Woodcock and Monte Davis, Catastrophe Theory, (Harmondsworth: Penguin Books 1980).

C.H. Waddington made initial efforts at describing these higher principles in his seminal work on epigenetic space.[8] Waddington suggested that mediating the gene complex and the phenotype there is an inner space of interactions, an epigenetic space, where 'a whole lot of controlling actions, each of which is brought about by network effects, or feed-backs'[9] play off against each other, creating a landscape of forces that canalise processes in the right developmental direction. As he put it:

> When an egg is developing, different parts of it will follow different courses of development, and eventually finish up forming different parts of the final animal: some parts becoming muscle, some becoming nerve, and so on. This can be pictorially represented in terms of an 'epigenetic landscape' in which when the process starts there is a single valley, but that later branches into two or more and these branches split up again and again, until they have formed a number of separated valleys corresponding to the separate parts of the adult animal.[10]

The notion of a single valley splitting up into many valleys is really a representation of the biases that affect the way the developmental process unfolds. As the embryo develops, controlling interactions that were at first of minor importance may become more significant and vice versa. Eventually the system has altered so much that its controls can no longer ensure the stability of the former

---

8 Waddington's original ideas were expounded in his Strategy of the Genes, (London: Allen and Unwin, 1957), and discussed for non-technical audiences in The Nature of Mind and The Development of Mind; Gifford lectures for 1971/2 and 1972/3, (Edinburgh: Edinburgh University Press, 1972 and 1973). Piaget discusses Waddington's notion in Biology and Knowledge, (University of Chicago Press, 1971), pp.10-25.

9 C.H. Waddington, Tools for Thought, (St. Alban's: Paladin, 1977), p.111.

10 ibid. p.109.

pathway. It may then break down into a general chaotic turmoil, or, as most often happens, it undergoes a branching into two new paths each with its own stability.

Now it seems possible that knowledge describes certain structural features of the landscape of branching valleys in cognitive or neural processes. According to our knowledge-level theories the neural processes underpinning cognition are so organised that at certain moments they can be said to be invoking, exploiting, or manifesting different knowledge states. Viewed at a neurophysiological level there may be nothing characteristic in these transitions of knowledge states. In semantic processing, for instance, there may be nothing characteristic about the neural processes which 'realise' the invocation of different bits of semantic knowledge. If a system knows the meaning of 'red', for example, it has a concept of red, and in some way accesses its knowledge of that concept in the process of understanding sentences which contain the word. Knowledge of other concepts are called upon in the same way. The entire process of calling up knowledge is one we know nothing about. It is fairly clear that since the brain is a massive parallel device, its invocation of information is not achieved, as in serial machines, by calling on specific addresses. Knowledge of the meaning of terms is bound to be distributed throughout the cortex and therefore accessed through some distributive process.[11] But there need be nothing telltale about the

---

11 For attempts at modelling some of these processes see Geoffrey Hinton and James Anderson (Eds.) Parallel Models of Associative Memory, (Hillsdale NJ: Lawrence Erlbaum, 1981), particularly the articles by Hinton "Implementing Semantic Networks in Parallel Hardware", and Anderson and Mozer "Categorization and Selective

neural activity that constitutes accessing such knowledge. From a study of neural activity alone no one might identify the beginning, middle and ending of the process of accessing semantic information. And yet at a higher level of analysis that is exactly what is happening. The advantage of representing this complex process of accessing as a landscape with branching valleys -- 'attractors' in the language of catastrophy theory -- is that we may find a mathematical account of the shape of these valleys which sheds light on patterns of activity at the neural level. The neural processes themselves change gradually, incrementally. The outcome of such gradual change is a sudden jump from a state of ignorance about the meaning of a sentence to understanding it. Along the way there are plateau points, stages of achievement that also represent jumps to new states of knowledge. If the dynamics of such transitions are describable mathematically, as in catastrophe theory, new types of experiments may be suggested. The long bridge between neural processes, cognitive mechanisms and knowledge states may be narrowed.

I grant that the three analogies I have offered are so schematic as to be virtually worthless for empirical research. The plain fact of the matter is that no one has any clear idea ofhow knowledge states constrain cognitive and neural processes. And no one has any evidence that they do. The connection between the two or three domains represents uncharted territory. Nonetheless, for all their unclarity it is analogies of this type that I have in mind when I think of

---

Neurons"; and Hinton's more recent work with Terrence J. Sejnowski "Optimal Perceptual Inference" to appear in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

knowledge constraining mechanism. Nowhere do I think of knowledge states actually being encoded declaratively in data structures or embedded procedurally in rules. In my view, it is sufficient that processes unfold in the appropriate manner. Thus both parallel processes and analogue processes might instantiate knowledge. It is certainly not necessary that the system be computational in the Von Neumann sense.[12] Nonetheless, the very idea that there might be a bridge of some sort between knowledge level analyses and analyses at lower levels presupposes that the organisms genuinely have states of knowledge. It is idle to expect knowledge analyses to teach us anything about constraints on processing if the organism doesn't satisfy these knowledge specifications. If knowledge were a mere artifact of theory, a computational convenience, as some would say, the knowledge landscape would not be like the epigenetic landscape; it would be a mere construct, divorced from the causal processes that guarantee its descriptive adequacy. In short, it would be a fiction.

The majority of attacks on knowledge level research, then, come from critics who are skeptical of the reality of knowledge. I have a measure of sympathy with certain of these complaints as I shall explain in the following two chapters. But most are based on pure misunderstandings of what realism about knowledge commits us to. The fault does not lie with the critics alone. They have quite naturally

---

12 In most of today's computers, data and instructions are passed back and forth sequentially through a single link to a central memory. The basic architecture was conceived by Von Neumann and forces computations to proceed serially, even when parts of a program could be handled in parallel. See John Backus, "Can Programming Be Liberated from the Von Neumann Style?", Communications of ACM, Aug.1978.

directed their criticisms against a view of realism that is, "The only one on the market" so far. This is the view which Chomsky and Fodor have elaborated. By contrast the realism I have been endorsing is more abstruse. It can handle their complaints.

In the remainder of this chapter I shall review some of the arguments against realism. I begin by presenting the Chomsky-Fodor realism, then discuss what seems to me its strongest challenge. I argue that directed against the narrow realism they propose the criticism applies. But a realism of a more liberal sort might be immune to standard criticisms.

Representational Realism

Chomsky and Fodor both defend the view that epistemic states -- that is, knowledge, belief and their cognates -- are real states of a system only if they are **encoded** in an inner language. Chomsky has been less forthright on this point than Fodor. Nonetheless he has, on several occasions, made it quite clear that a grammar is known by a person in the way a program, or a set of data structures, are in a computer.[13] Fodor's formulation is more specific. According to Fodor, a person knows or believes some proposition p **if** he stands in a certain 'functional' relation, a believing or knowing[14] relation,

---

13 For instance in a reply to Harman, Chomsky wrote "Again it is possible to design a computer that generates in this manner (and in fact, there has been a fair amount of experimentation with such programs)." In S.Hook, (ed.), Language and Philosophy, 1969, p.156.

14 Fodor argues that attitudes such as believing, desiring, hoping, intending and so forth, can be identified with functional states of human cognitive systems. The content of these attitudes is identified with a further functional state, namely, the state associated with the functional role of the relevant sentence in the language of thought. Since knowledge is an achievement term, signifying that the proposition believed is true (justified, grounded etc.), it cannot be distinguished from belief, or justified belief, purely by

to a sentence S, in the language of thought (where S represents p). Thus for every proposition known or believed there will be an actual sentence in mentalese playing an appropriate role in the functional economy of the person. If the person knows Euclid's axioms, for instance, Euclid's axioms will appear (in suitably translated form of course) in some part of the person's brain. If he knows some of the theorems that follow, they too will be encoded and located in his brain somewhere. By contrast, virtual or implicit beliefs, such as the class of Euclidian theorems the person has never thought of, are not encoded.

At the time it was presented, the Chomsky-Fodor brand of realism seemed particularly apt to rebut a standard objection levelled at knowledge level research. A favourite objection of those disbelieving the utility of knowledge level research is that there is no principled distinction between abilities that are cognitive, and which therefore are supposed to require knowledge of rules, principles or facts, and abilities that are non-cognitive, and which therefore presuppose no knowledge -- at least the sort of knowledge most naturally expressed in propositional form. If the cognitive/ non-cognitive distinction is ad hoc, then any ability whose manifestation displays a pattern describable by rules, that is, virtually any ability, can be described as the result of a body of knowledge, whether it be the ability to ride a bicycle, to swim or to walk down the street. Obviously such a move vitiates the plausibility of the entire knowledge level research

---

its functional role in the system. Knowledge makes unavoidable reference to the world, and hence stands outsides Fodor's methodolically solipsistic psychology. Thus for Fodor tacit knowledge is really tacit belief. This said, since Fodor himself does nonetheless talk of tacit knowldge, I shall continue to place him in the company of those with theories of tacit knowledge.

program. For if any ability, whatever its sophistication or its functional role, can be called cognitive then knowledge can be attributed to any system, from the most primitive onward. In that case, attributions of knowledge could set no constraints on lower level processes. For as the complexity of an ability falls, the diversity of systems able to display that ability explodes exponentially, decreasing the chances of discovering any design features which all or most such systems must have.

By distinguishing cognitive abilities from non-cognitive on the basis of whether, in the expressing of the ability, encodings of the rules or representations have an actual causal role to play, Chomsky and Fodor had an operationally precise definition of cognition. The difference between abilities that are merely described by rules, then, and abilities that are cognitive, is that cognitive abilities -- or rather the manifestation of cognitive abilities -- are driven by knowledge of rules, whereas non-cognitive abilities are not. In other words, cognitive performances involve causal chains that unavoidably involve mental or representational encodings of rules or propositions.

By identifying tacit knowledge with functionally active sentences encoded in the language of thought, Fodor and Chomsky had an easy account of why theories of tacit knowledge are explanatory theories and not merely descriptive. As Fodor noted in The Language of Thought.

> A by now chestnut of a question that is supposed to embarrass information flow psychologists goes like this: If you are willing to attribute regularities in the behaviour of organisms to rules that they unconsciously follow why don't you say (eg) that the planets 'follow' Kepler's laws in pursuit of their orbit round the sun? ...
>
> It should be clear how this sort of question is to be dealt with. What distinguishes what organisms do from what the planets do is that a representation of the rules they follow constitutes one of the causal determinants of

> their behavior. So far as we know, however, this is not true of the planets: At no point in a causal account of their turnings does one advert to a structure that encodes Kepler's laws and causes them to turn.[15]

According to Fodor tacit knowledge of rules genuinely explains behaviour only if those rules are actually represented in a cognitive system. Since cognitive behaviour is driven by rules, and non-cognitive behaviour is driven by biological forces or control mechanisms of another type, the rules guiding cognitive behaviour must be internal to the system. They must be 'embedded' or 'encoded' in the control structure of the system and must be invoked appropiately. By citing these encoded structures and the order of their invocation, Fodor contends, we can explain cognitive behaviour. For we will have a theoretical model of a system which simulates behaviour.

I think we may agree with Fodor and Chomsky that it is the causal rôle played by an agent's knowledge or belief about rules, which decides whether he is acting in a rule driven manner or simply in a way that can be described by rules. Whether this permits a general distinction between cognitive and non-cognitive abilities I would not like to say. But the question that interests us now is whether knowledge of rules must be encoded in some internal language-like system to be causally active. Both Fodor and Chomsky readily agree that tacit knowledge is unlike familiar knowledge. If a rule or fact is known, in the ordinary sense, the knowing subject must be aware of the rule or fact, or be able to immediately formulate it when prompted. Similarly if the subject 'implicitly' knows a rule he must be able to recognise and assent to the rule when offered a formulation of it. These are standard criteria of knowledge. One who knows rules

---

15 The Language of Thought, p.74.

tacitly, however, may not, in any circumstance, be in any better position to say what he tacitly knows than anyone else. Tacit knowledge has a unique functional role in the system. Although many philosophers have been led by these differences to question whether the notion of knowledge is appropriate here, Fodor and Chomsky maintain that if there is evidence that within cognitively competent agents there is to be found robust functional states whose control over behaviour and other internal states is commensurate with their content, we shall have good reason to assume tacit beliefs are real states. Moreover, this tacit knowledge will be encoded declaratively or perhaps procedurally in some inner notation. For given the syntactic-mirroring hypothesis we cannot expect a state to exercise causal powers commensurate to its content unless it includes a syntactic structure whose formal properties have identical causal powers.

In Chapter One I gave several reasons for supposing that knowledge, whether tacit or occurrent, need not be encoded declaratively or procedurally to be a causally active agency in the system. Nonetheless, like Chomsky and Fodor, I accept that knowledge states are real states of a cognitive system; they have causal power and can be identified through scientific inquiry. It is important, therefore, to sift through some of the standard arguments against representational realism to decide whether a different form of realism about knowledge might be tenable. In what follows I propose to examine some of these arguments to demonstrate that if we have a different conception of how knowledge might be causally active, we can remain realists about knowledge despite rejecting the doctrine that all knowledge in encoded.

The Intellectualist Fallacy

The first argument against realism of rules I wish to consider is the oldest. Ryle called it the intellectualist fallacy and would dutifully reproduce it whenever anyone argued that "the execution of intelligent performance entails the additional execution of intellectual operations."[16] Ryle thought it ludicrous that underlying our capacities to regulate action intelligently there must be recipes or programs, that we follow. If thoughtful scientists have recourse to such notions, he suggested, it is to describe structures implicit in performance, not to describe structures in underlying process. His argument took the form of a reductio ad absurdum, revealing the vicious regress latent in 'intellectualist' reasoning.

A facsimile of this argument might run like this. Rules unaccompanied by methods of interpretation are meaningless, blind. Hence, if a system's cognitive competence depends on knowledge of a set of rules, the system must also have knowledge of how to apply those rules. A rule enjoins action only to an agent able to grasp its import. Yet surely grasping the import of a rule is itself a cognitive process. It requires knowing what the rule means and so is itself a cognitive ability. Since, ex hypothesis a system only has cognitive ability in virtue of its (tacitly) knowing a set of rules, there must be a second set of rules tacitly known: the rules telling the system how to interpret the first set. These in turn will require another set instructing it how to interpret the second set and so on ad infinitum. An infinite regress is generated and nothing explained. Only if it is

---

16 The Concept of Mind, (Harmondsworth: Penguin, 1970) p.48.

assumed that a rule driven system has built into it the power to interpret rules simpliciter, can it be said to actually implement those rules.

But if the system has a **primitive** ability to interpret rules, why not a primitive ability to perform the cognitive task itself? Is it not arbitrary to say that first order rules must be known (encoded) but second order rules must be built in? Why not simply say that the reason a system behaves in conformity with a set of rules is that it was designed to do so in the first place? At the very least, extra grounds are needed to establish that the system encodes anything. For encoding cannot be necessary for cognitive capacity since some cognitive capacities must be primitive. Consequently, the touchstone of a cognitive process cannot be that its processes are driven by encoded rules.

This, then, is the first major attack on representational realism. It cannot be said to prove either that rules are not encoded in a system with cognitive capacity or that they are not tacitly known, for it shows only that:

(1) a system need not have rules encoded anywhere in its innards for it to display behaviour that is effectively described or specified by a set of rules; and

(2) it cannot have rules encoded for all its structural activities, cognitive or otherwise.

But although not decisive the argument does seem to shift the onus of proof to those claiming to be realists about tacit knowledge. The intellectualist fallacy establishes that it is possible to describe a

body of knowledge whose explicit possession by a literate, computationally competent agent would endow him with the ability to perform a given task, without maintaining that flesh and blood agents, whose native ability is matched by hypothetical rule driven computational beasts, really have such a body of knowledge as causally active states in their minds or brains. It therefore allows us to view tacit knowledge as an oblique characterisation of an ability: a characterisation which does not say anything particular about the structures and processes actually occurring in the agent. Just because the patterns discovered in behaviour are most easily described in terms of a set of rules, we cannot infer that the cause of these patterns lies in the agent's knowledge of these rules. For first, the agent may not in any sense 'know' the rules. And second, even if the agent does know the rules, it is not necessary that he rely on that knowledge to act intelligently. Other control processes may be at play. The regress argument does not, of course, show that in fact knowledge is not causally active. A system may, in fact, organize its behaviour because it knows certain rules. Similarly, a system may, in fact, encode that knowledge in some inner language-like system and rely on control structures similar to those found in serial computers. All that follows from the regress argument is that it cannot be necessary that a system or person know a set of rules to be truly said to have a cognitive ability.

A Rejoinder

Now against this, adherents of representational realism have a strong rejoinder. Ryle's argument, they contend, applies to behavioural

abilities only; it shows that tacit knowledge may not be required to explain why agents are able to regulate their behaviour in an orderly, apparently rule governed fashion. But in the case of cognitive abilities, particularly language abilities, part of what must be explained is orderliness in acquiring occurrent knowledge. As Michael Dummett has argued:

> A good deal of conscious knowledge is required for the knowledge of language. Someone may be in doubt whether he has interpreted a sentence correctly, but he can seldom be unsure whether he has put any interpretation on it at all: yet if understanding were simply a practical ability -- say to respond appropriately -- there would be no reason why he should be in a position to say whether or not he has understood.[17]

Dummett's point is that the occurrent knowledge which arises every time we exercise our linguistic competence is an essential feature of that competence. It is obvious that we do have such occurrent knowledge, for if we did not, then on those occasions where we fail to react behaviourally to something someone has said, we would not be in any better position then anyone else to say whether we understood what was said. But of course we are. We may be mistaken in thinking we have understood someone correctly. But we cannot be mistaken about thinking that we have put at least some interpretation on his remarks. Understanding characteristically involves conscious[18] (or better occurrent) knowledge.

---

17 London Review of Books, 3-16, Sept.1981, p.6.

18 Although Dummett uses the expression "conscious knowledge" in a non-technical sense, it may be wiser to abjure references to conscious knowledge when speaking of understanding, and identify the temporal immediacy of understanding instead with something like occurrent knowledge. Cf. Wittgenstein, Philosophical Investigations, secs. 152, 153 et passim.

Moreover, this knowledge is partly constitutive of linguistic ability. People must know what they are talking about. They must occurrently know what their words refer to. To speak about some object or to understand what someone else has just said about some object, we must occurrently think about the object. We shift from occurrent knowledge state to occurrent knowledge state in the course of conversation. A theorist cannot state what people do when speaking or understanding without stating what occurrent knowledge they are transmitting or receiving. To give a theoretical description of what it is to speak and understand a language, then, we must describe what a person would mean and understand by the various sentences he could utter and comprehend. By contrast, to give a theoretical description of a practical ability we need only describe behaviour. We have no need to mention knowledge, in either its occurrent or tacit forms, because we can state what it is to master a practical ability without stating what a person must occurrently know to participate in that practice.

Furthermore, the only reasonable explanation of how we come to gain occurrent knowledge, treats the event of coming-to-know as the outcome of mental acts like unconscious judging or unconscious inferring. Since, presumably, these require knowledge of concepts, and knowledge of rules of inference, we may infer that knowledge of rules and concepts are real states of the system that are exercised whenever we come to know something. This more or less is the unwritten claim in Dummett's defence of realism about tacit knowledge. It applies mutatis mutandis to all cognitive abilities.

Such an argument is sufficient to show that the intellectualist fallacy as it stands does not apply to cognitive abilities. Ryle may have been right in arguing that most practical abilities do not, in general, have a genuine epistemic component. But there are disanalogies between practical and cognitive abilities which rule out the standard intellectualist argument. Further arguments are needed to show that cognitive abilities are not dependent on tacit knowledge of rules and axioms.

Another Argument

Such arguments are not long in coming. An agent who unconsciously infers things, as allegedly he does during perception, speaking, understanding, and all other cognitive activities, obviously does not have knowledge of his rules of inference in the sense of being able to state them when asked or of being able to recognize them when formulated for him. His knowledge is tacit. This point has been stressed too often to need repeating. Tacit knowledge lies beyond the introspective horizon, outside the reach of expressive faculties. But this extension of the normal sense of knowledge requires defence. There are clearly numberless beliefs we have which we cannot dredge up introspectively. Self-deception, wishful thinking, repression and all the rest are common sense explanatory devices we invoke to sanction the reality of unconscious belief. What differentiates tacit knowldge is that we may be conceptually unqualified to understand the propositions they allegedly represent. Knowledge qua intentional state is a relation the knowing subject bears to a (true) proposition, state of affairs, fact, or what have you. It is a precondition on bearing this

relation that the proposition be graspable, that the knowing subject knows what it is for that proposition to be rather than not be. He must know the truth conditions, or assertability or verifiability conditions for the proposition or state of affairs known. In short, the subject must understand what the proposition means. To understand, however, he must have the appropriate concepts. For propositions are grasped through grasping the meaning of the concepts involved.

But what evidence is there that the subject has these concepts? It is a condition on knowing a concept that the agent who knows the concept is not totally restricted in his use of the concept. If I have a concept of F, then I can think that a is F, or b is F or c is F and so on, for all individuals that do not presuppose not F. I am not restricted in my use of F to just one individual. Knowing a concept is an ability I possess that can in principle be manifested and expressed in diverse ways. Similarly, if I have a concept of some individual, a, then I can think that a is F, or a is G, or a is H. Knowing an individual is knowing it as a bearer of attributes: not just a single attribute, but potentially many. Gareth Evans[19] has called this the Generality Constraint and has made it the touchstone of thought and representation: any system of thought must conform to the Generality Constraint. The problem with tacit knowlege is that the concepts which are involved in tacit rules are so restricted in application that they fail the generality constraint. The agent can display his knowledge of those concepts in so few ways that we may doubt whether he really knows those concepts. We doubt whether he understands, even tacitly, what

---

19 op.cit. sec. 4-3, et passim.

those rules mean. Yet if a subject does not understand the rules he allegedly knows, in what sense can he be said to really know them? If a rule exploits four concepts, but each one has a use in that rule alone, how can we show that the agent understands those concepts? The concepts fail the test of the generality constraint.

This is the core of the second argument against realism: agents do not genuinely know tacit rules, for they lack the requisite conceptual abilities.

In one sense, the generality constraint is simply a test of the robustness of the concepts attributed to an agent. I shall discuss this notion more fully in a moment, but the basic idea is straight forward. States, properties, processes, etc., are robust to the extent that they lie at the intersection of analyses. If attributing an agent knowledge of a concept would explain his behaviour in several apparently unrelated contexts, then there is strong evidence the agent has that concept. Thus, if, on the basis of an isolated action we would consider ascribing an agent the belief that Aristotle was a philosopher, we will do well to consider what other actions of his suggest that he has the concepts of Aristotle and philosopher. If no other actions of his are best rationalized by attributing him beliefs about Aristotle and about philosophers, then we will be hard pressed to justify our initial interpretations of his actions. We should cast about for other interpretations of his actions.

Now it is a characteristic feature of tacitly known rules that they have no connection with the thinking, evaluating and reasoning systems. Indeed, rules are tacit precisely because they are in principle inaccessible; agents can neither reason <u>with</u> them, nor reason

about them. The only place such knowledge states can be put to use -- that is, manifested -- is in the specified faculty for which they were postulated, as is the case, for example, in Transformational Grammar, where tacit knowledge of transformational rules are attributed to speakers who lack explicit knowledge of the requisite concepts to understand the rules, and where they can be said to know how to apply the rules but not to know what they mean.

What shall we say of such apparently unrobust knowledge? Shall we stick to our familiar model of explaining occurrent knowledge as the result of inference, this time, however, postulating an unfamiliar type of knowledge -- tacit knowledge -- as the premises in the syllogism? Or shall we rather save ourselves postulating strange knowledge states by adopting a new paradigm of explanation, one which does not require us to see occurrent knowledge as the outcome of inference but which will account for the acquisition of occurrent knowledge, by more robust means? The question may seem moot: having no other explanation available we appear to have no choice but to appeal to tacit knowledge. But accepting the intentional framework for the time being does not commit us to realism about the tacit knowledge it posits. One can assume a more pragmatic, anti-realist stance to tacit knowledge.

Thus theories of tacit knowledge might be interpreted in the way kinematic theories in physics are. They serve as non-causal explanation. For instance, we can describe the trajectory of an object as the sum of force vectors along the X, Y and Z axes. These forces allow us to predict the spatial course of the object through time. See Fig. 3.4. They do not, however, tell us anything about the processes or real forces that propel the object. Force vectors are constructs

introduced to facilitate computation, artifacts of our formalism. The force that propels the object has only one component; it is not literally the sum of three distinct forces.

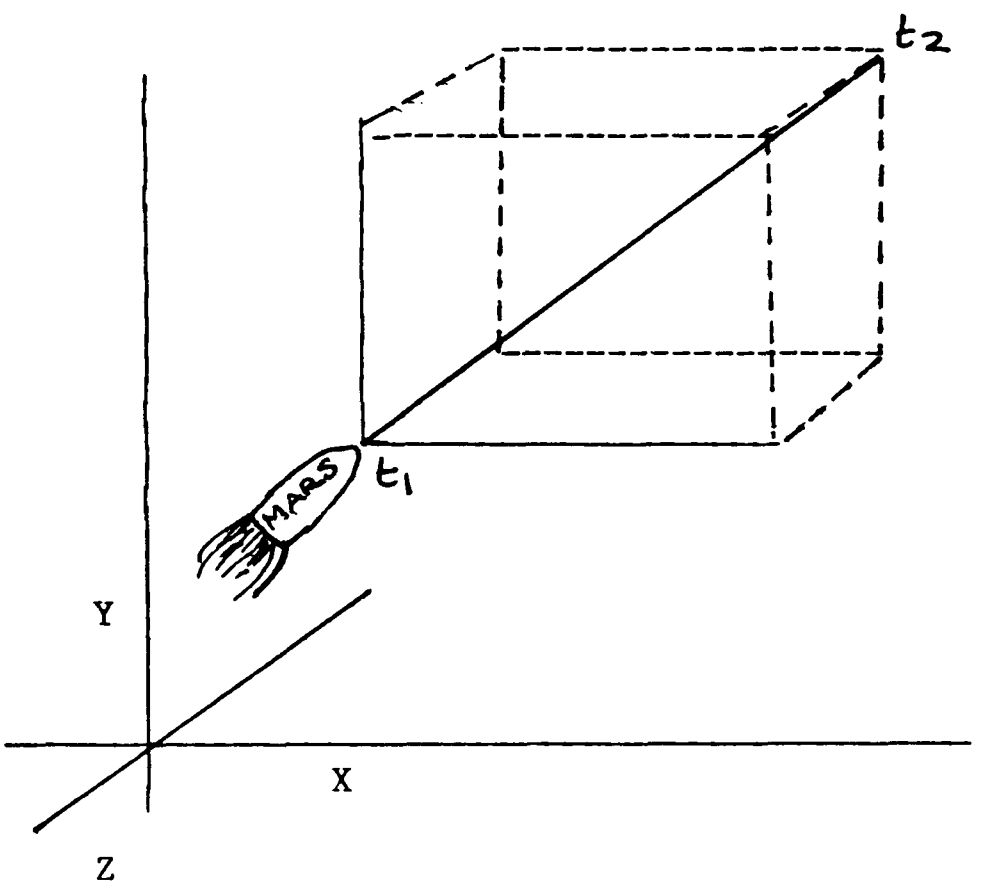


Figure 3.4

Similarly in biology D'Arcy Thompson[20] has elegantly demonstrated how many of the patterns in nature are really more the result of shaping influences stemming from the environment than the result of influences internal to the system. Nature has an overbearing concern for economy, efficiency and parsimony. Water follows the easiest path to sea-level, objects under pressure buckle at points of minimum strength. In general, patterns and shapes come into being in forms that reflect these conservative concerns. Strictly speaking, the range of forms that are possible for a system are a function of <u>both</u> the

---

20 <u>On Growth and Form</u>. (Cambridge: Cambridge Univ. Press, 1917).

internal properties of the system and the environmental forces to which it will be subjected. But we can often predict the final form of a system without concerning ourselves with either its internal properties or the mechanisms by which inner and outer 'forces' equilibrate. To return to evolution, from a study of an organism's niche we may be able to predict the morphology of that organism given only the most minimal information about its internal constitution. For the external forces are the decisive ones. They shape its form and behaviour. No one would suggest that these accounts are not explanatory. They allow us to predict form and behaviour from function. Moreover they have an ultimate causal basis in the mechanism of natural selection. They are legitimate scientific explanations. Nonetheless, predictions are made without concern for internal processes. They serve as non-causal explanations.

Might not the appeal to tacit knowledge in psychological explanation be no different than the appeal to force vectors in kinematics, or to conservative principles in biology? Why not view tacit knowledge as an artifact of intentional explanation, a construct which allows us to predict, and in a certain sense, explain the acquisition of occurrent knowledge without identifying causal processes internal to the system? Without grounds for thinking that tacit knowledge is robust, there is no reason to see tacit knowledge as a causally real state of a system.

This is the full form of the second argument realists must overcome: They must show that tacit knowledge of rules and the concepts they presuppose are robust states of the organism. They are genuine states.

I want to show now that this argument is only compelling against the brand of realism that identifies tacit knowledge with functionally active sentences in the language of thought. As with the intellectualist argument discussed earlier, it shifts the burden of proof from anti-realists to realists -- those who believe that states of tacit knowledge are robust states of systems. Since knee jerk realism is clearly out of line in many of the theories physicists and biologists use, we need a stronger reason to entify states than the bare fact that they figure in the best theories we have so far. I intend to show that there is reasonable evidence that states of tacit knowledge are indeed robust -- they satisfy a restricted version of the generality condition. But this evidence does not support the further claim that such states 'include' sentences encoded in the language of thought. The representational realist believes both that cognitive competence depends on tacit knowledge states and that these states are encoded. We may separate the claims and consider the evidence for each. Let me begin with a discussion of the concept of robustness.

When is Tacit Knowledge Robust?

Robustness is a technical term introduced by statisticians and methodologists to designate the degree of reality which a construct has. According to Donald Campbell[21], a construct is robust if it can be measured by a number of independently derived indices and if the

---

21 "Common Fate, Similarity and Other Indices of the Status of Aggregates of Person Social Entities", Behavioural Science, 1950, pp.14-25; and Donald Campbell & D.W. Fiske "Convergent and Discriminant Validation of the Multi-trait Multi-matrix Method", Psychological Bulletin, 56,(1959), pp.81-105.

various indices order the same measured objects in the same way. The more indices we can find which concur, the stronger our confidence that we are measuring something real. For presumably if an object or property exists it can be multiply identified. It can be registered from different points of view and so located through 'triangulation'.

The concept of robustness has played an important role historically. For instance, the distinction between primary and secondary qualities is essentially one between qualities which can be cross-modally validated directly and ones which cannot. Shapes are real, causally efficacious features of the world, colours are not. The one inheres in objects as a relatively stable property. The other is receiver-relative: change the receiver and the colour changes. Similarly, the distinction between objective and subjective rests on a difference in degree of detectability -- objective properties are those which can be detected by a variety of means and subjects.

Now if tacit knowledge of rules and concepts can be identified with states that have some claim to reality, the states so identified should be discernible from different points of view. They should be capable of manifesting under different conditions. For if a state is real it should be capable, in principle, of participating in more than one causal chain. This was the point of the generality condition. Thus if we were to alter the standing conditions, or the background conditions of the system -- say some of the other desires and beliefs of the system -- we would expect a change in the causal power of the state.

It may prove helpful if we compare the status of tacit knowledge with that of modules or work stations in a factory. Imagine that we

have been given the job of discovering the design of a factory. We know the inputs to the factory, or at least we have a description of its input, and we know its output, that is, we have a description of its output that is 'consonant' with our description of input in the sense of stemming from the same interpretation manual. Also, we believe we know the kinds of functions that are natural for the functional units, modules, work stations etc, of the factory production line. We can then proceed to analyse the way modules might be organised on the production line, given a few further assumptions such as that each module performs only one job, the jobs are performed serially, in stepwise fashion, and so forth. For the sake of our analogy, our problem now is to decide whether our breakdown of the production line into modules identifies modules that are robust. Do the modules postulated perform real jobs within the system, or are they mere constructions, fictional units that fail to denote anything real in the system?

In standard scientific theories one tests the validity of constructs by assessing the theory's truth. Since theory is true in virtue of the reality of the conditions it describes, each referential term in a theory must denote something in the world. Those very conditions constitute the truth conditions of theoretical statements.[22] If a theory is not only well-confirmed but, among other things, extendible to new domains of phenomena, compatible with other confirmed theories, capable of subsuming or being subsumed by other theories

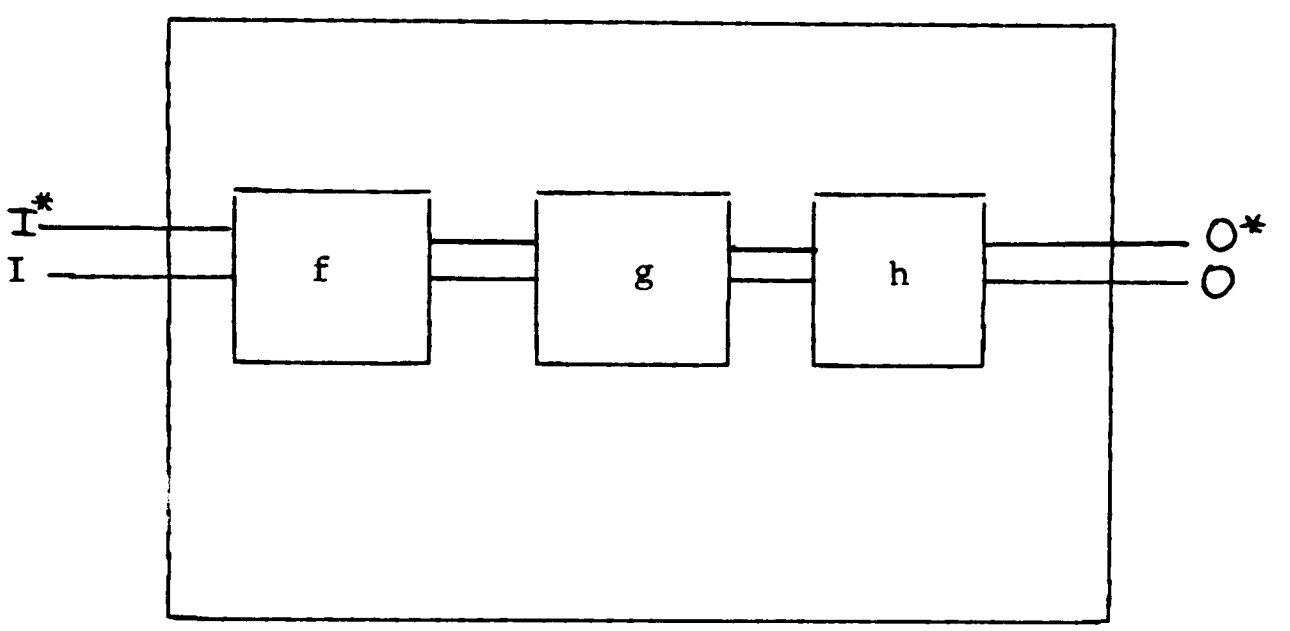
---

22 This is no more than a restatement of Tarski's theory. It applies whether we deem the reference of predicates to be properties or sets of individuals.

(when suitably supplemented by bridge laws), simple and elegant, then it has strong claims to denote robust entities. With functional decomposition our methods of testing robustness are a bit more recondite, but substantially the same. Once a decomposition has been shown to be empirically adequate, that is, once its parts are shown to interact in a way that produces the known input/output table of the system, we then question its simplicity, elegance, and potential extendibility. And, if the system is organic we evaluate a decomposition further by comparing it with accepted decompositions of related creatures. The existence of a heart and liver in all mammals is a good indication of the existence of a heart and liver in man. We also study ontogenetic development, that is, the changes the system undergoes as it matures/learns its skills; and we can examine the effects of disease, injury and damage.

In our production line case, most emphasis is laid on potential extendibility and simplicity, though if there were evidence of the effects of damage to the system that too would lend weight to a decomposition. Since a functional unit is defined by its role in producing a particular input/output schedule we cannot secure confirmation of its robustness merely by exposing it to inputs that are already listed in the canonical input/output schedule. Such exposures confirm that the defined units are in fact empirically adequate. But to test for robustness we want to establish more: we want to show that each part is a relatively separate entity with its own dispositions. It is not just a relationally defined state; it is an actual part of the system, it has a categorical basis. The simplest method of proving robustness would be to interfere with the working of individual parts

and study the results. But failing our ability to damage or suppress internal parts we must continue the investigation behaviourally. Hence we will need to move beyond the confines of the canonical table to input conditions that are not involved in defining the unit, and witness the results. See Fig. 3.5



I,O = normal input and output
I*,O* = extraordinary input and output

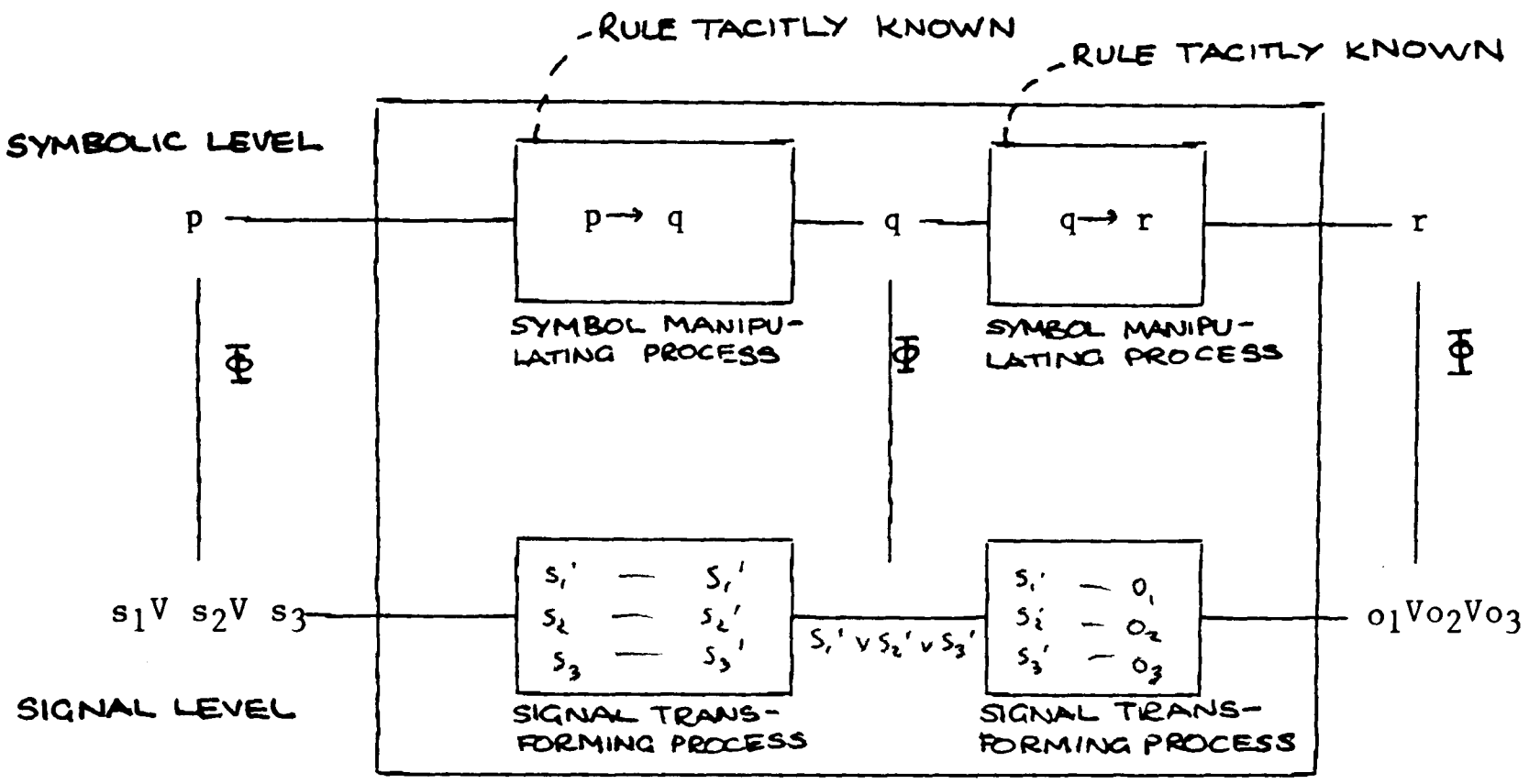
Figure 3.5

For example, if a system customarily receives input at the rate of 5-10 units per sec., we may try changing the frequency and duration of input, altering the flow of electricity, gas, widgets and other resources utilised as input by the system. As the system begins to overload or starve, different components may be expected to react differently. The effects of input changes are not usually constant throughout the system. Thus, as we build up evidence of behaviour under abnormal conditions we get insight into the potential of each part.

Essentially the same approach is used in psychology. A common ploy to test functional models of psychological processes is to present

a subject with several signals simultaneously, then watch to see how it selectively reacts to some and not others. Or again, if dozens of sentences or sentence-parts are flashed at a subject for brief periods, its response will often be ridden with errors, displaying underlying dispositions, patterns of coping, and reactions to stress which can be revealing about the internal make-up of the system.

Now Fodor and Chomsky treat rules that are tacitly known, as being functional components of a system. They are not identical with standard functional components, however, for they are identified in intentional language: they are defined over representations. Thus, the input states they transform are not mere signals that are measurable numerically, or otherwise describable in the language of standard science; they are symbols; their identity depends on what they mean, what they represent. Accordingly, although Chomsky and Fodor see these rather special components in the system as serving an input-output function, this function cannot be defined in the normal way. It presupposes that the system endowed with rules can treat signals as if they are contentful. It can interpret them. Figuratively we can represent this mutant functionalism as in Fig. 3.6. Due to a system's interpretative capacity it treats a class of otherwise structurally dissimilar signals as representationally equivalent, mapping them into another representationally equivalent class of signals, so that at the content level a well-defined operation has been performed.



Φ = interpretation function that allows cognitive scientists to interpret the function of a signal transforming process as a symbol manipulating process.

Figure 3.6

Now if we could access these symbolic rules introspectively, the rules would exist as entities for the introspective system. That would provide a measure of reinforcement for our linguistic theory. For the rules would then lie at the intersection of two causal pathways: the pathway leading to linguistic performance, and the pathway leading to consciousness. Since we cannot access rules introspectively, however, their robustness must be established in more behavioural ways. Yet Chomsky has denied the direct relevance of psycho-linguistic testing of tacit rules. His grounds are that states of tacit knowledge are 'embedded' in cognitive machinery in non-simple ways. They may be there, perfectly 'encoded' in syntactic structures, but the processor which uses them is somewhat impotent and restricted in its access to scarce cognitive resources. Alternatively, they may be there, again created in some inner language but so deep within the cognitive

system that behavioural studies cannot get inside the system to test rule-use experimentally, by means of techniques such as reaction time etc. In both cases we are told behaviour is an unreliable measure of competence. See diagrams in Introduction.

Accordingly, having forsaken psycholinguistics as the means of proving robustness, Chomsky is obliged to adopt the method of studying the behaviour of systems outside their normal input/output schedule. If he can show that linguistic theory explains performance in non-standard conditions, he can obtain further confirmation of the robustness of rules. And in fact, studies have shown that if native speakers are asked to speak poetically, or to invent new locutions, or to violate category restrictions, there is method in their deviance. The new locutions bear a definable relation to standard grammatical categories and are combined in ways that preserve deep grammatical rules.[23] Extensions are compatible or at least continuous with existing rules. Speakers seem to exploit their knowledge to create new structures that preserve principles. Apparently, old knowledge serves as a robust core for new skills.

Nor is evidence restricted to deviant performance. Additional evidence of a similar kind could be had by studying patterns in the way competence grows or decays. If a competence is composed of a set of semi-autonomous rules, some rules can be gained or lost without affecting the remainder. If we assume that corresponding to each rule system G there is a class of linguistic structures, L, which can in principle be generated, then as we add or subtract rules from G, L should vary in a predictable fashion.

---

23 James Higginbotham, personal correspondence.

In point of fact this last approach most clearly reflects Chomsky's own efforts. Since learning is a case where rules are added, deleted or altered, studies of the dynamics of change in a speaker's grammar may cast light on the elements that are robust. For instance, learning is an excellent way of showing that the rules tacitly known are structured; they are operations that apply to highly structured elements: namely, a small vocabulary of phrase-markers. If a transformation rule is a mapping of labeled bracketings onto labeled bracketings,[24] the fact that the same labeled bracketings (phrase markers) begin to emerge as the dominant element in the transformation rules as a child moves closer to its steady state grammar, is evidence that the rules possess a combinatorial structure composed of robust parts. Such an hypothesis might seem reasonable on synchronic analyses alone. In any transformational grammer the rules cited partake of a small storehold of technical linguistic concepts such as verb phrase, noun phrase, and so on. But as the number of rules in a grammar is small, it is not impossible that each rule is treated as conceptually independent from every other. Accordingly, a speaker might have one 'meaning' for VP in one rule, but have other meanings for it in the rest. To disprove this hypothesis, attention must be paid to the trajectory of grammars a speaker/hearer lives through on its path to its steady state grammar. As the most plausible path takes phrase markers to be univocal across rules within any grammar, we have further corroboration that phrase markers are robust. Since rules are composed of phrase markers (among other things), the robustness of rules flows from the robustness of their parts.

---

24 "Linguistics and Philosophy" reprinted in Chomsky's Language and Mind, (New York: Harcourt Brace, and World, 1972).

All that remains to be done to defend the legitimacy of tacit knowledge of rules, then, is to show that the technical concepts making them up are understood. It will be recalled that we originally appealed to the notion of robustness to overcome the contention that tacit rules are not understood. They are not understood, so ran the argument, because we lack concepts of the technical notions presupposed by them. It is true, we cannot actually prove we lack those concepts, for there are too few contexts to display ignorance or knowledge of them. But the onus of proof lies with those who claim knowledge. For it is constitutive of knowing a concept that that one be able to manifest that knowledge in countless subtle ways.

Attention to the dynamics of learning may help in this matter too. Chomsky's ultimate defence of tacit knowledge rests on a theory of language learning. In what follows I shall argue that when theories of language learning are put in the proper perspective they do support attributing tacit knowledge to speakers. They do not, however, support any further claims about those concepts being identical with elements encoded in a language of thought. We can defend the robustness of knowledge level constructs without invoking the language of thought hypothesis.

Acquiring Tacit Knowledge of Technical Concepts

To set the stage for our discussion of learning tacit concepts let us be clear about one thing. There is no *a priori* argument against faculties of the mind having their own specific concepts. It is analytic that possession of a concept is an ability that must in principle be manifestable, but it need not be manifested in the behaviour of the system as a whole. If we can show that a faculty can,

for instance, 'recognise' internal entities, it is no fatal weakness that the output of the faculty has no direct counterpart in the behaviour of the system as whole. For example, large corporations are treated by the law as individuals. They have certain rights and obligations. The motive for treating them anthropomorphically is that they behave as rational agents; they have desires, beliefs, plans, they acquire information about their environment of activity and they undertake actions that are subject to legal interpretation. Accordingly, corporations 'think', they deploy concepts in their reasoning. In attributing corporations concepts we rely on an interpretation of their behaviour. Needless to say, the concepts which the corporation has, need not be identical with the concepts its human administrators have. To decide which concepts are known by administrators, we look to their local task environments. These local tasks are distinct from the corporate level tasks. They are defined internally. It is sufficient, therefore, that we can interpret administrators as rational in their reactions to their own environments. The argument applies mutatis mutandis to cognitive faculties.

Now what evidence is there that the linguistic faculty has sufficent rationality to be said to know certain concepts. It will be hard to utilize the standard apparatus of attitude interpretation in order to attribute concepts to the faculty for its desires are, ex hypothesis, restricted. If it makes sense to say the linguistic faculty desires anything at all, its desires are to discover the meaning of utterances, and to properly express thoughts in words. To achieve these desires it exploits a system of beliefs about the

structure and meaning of language. These beliefs are most effectively represented as a set of rules, so that the agent is said to know a set of rules about his language. But the linguistic faculty is restricted to linguistic contexts as the sole domain in which it can display its rationality. Naturally if we know of other spheres of activity where the agent could harness those beliefs in the service of his desires we would have further confirmation of the robustness of his beliefs. The credibility of attitude attribution varies directly with the range of actions effectively rationalized. It is for such reasons that evidence of efforts to speak poetically, to violate category restrictions, and so forth, add to the robustness of a certain set of grammatical rules. Yet failing such extensions of the theory, evidence of rationality can still be found in the dynamics of belief change. If the linguistic faculty changes its belief system in accordance with broad principles of rationality, if, for instance, it manages the totality of its linguistic beliefs to maintain regulative ideals such as consistency, comprehensiveness, and simplicity, there is further proof that the faculty is cognitive, exploiting processes analogous to those found in the reasoning system.

Unfortunately Chomsky's account of language learning has changed in the last 10 years. At present, he is more sympathetic to an overtly maturational theory:

> In some domains -- acquisition of language, object perception, etc. -- the growth of our knowledge just happens to us, in effect. The mental grows from its initial to its steady state without choice.... In other domains -- the natural sciences, for example -- the growth of knowledge involves deliberate inquiry involving hypothesis formation and confirmation, guided no doubt by

> "abductive" constraints on potential hypotheses as well as other equally obscure factors that enter into choice of idealization and the like.[25]

This seriously complicates efforts to show that the dynamics of belief change can be used as evidence of the robustness of tacit linguistic concepts. For if maturation implies that there are no epistemological principles causally at work in language acquisition, it is hard to see how the trajectory of grammars which the language learner passes through en route to its steady state grammar, could bear on the robustness of linguistic concepts. Of course, it is not obvious that maturation does imply the absence of tacit knowledge of epistemological principles. It certainly does not, as I have been construing tacit knowledge. For according to the principles characteristic of the knowledge level, a language acquisition device could tacitly obey epistemo-logical principles without having encoded those principles anywhere. Changes in the linguistic faculty might just unfold in accordance with those principles in the way cells differentiate in accordance with homeorhetic principles. Thus, although questions of maturation versus learning raise problems that might have a bearing on the robustness of tacit concepts, I propose to bracket the question and proceed on the basis of Chomsky's old analysis. Accordingly, if we consider language acquisition itself to be an overtly computational process, our question is this: How can we rely on learning to justify the reality of tacit knowledge of technical linguistic concepts?

---

25 Rules and Representations, pp139-40.

In effect, the answer has already been given. The disagreement is over its interpretation. If the trajectory of grammars being constructed in the early stages of language learning can be seen as obeying basic epistemological principles, whereby each successive grammar appears as a rational transformation of a previous grammar, then a grammar may, with some justification, be thought of as a theory of a particular language. It will appear to consist of a system of hypotheses that lie at the "end of long and intricate chains of quasi-inferential steps."[26] We are justified in regarding such hypotheses as states of knowledge then, not because they can be manifested in 'countless subtle ways',but because they have been acquired in the right way. A state counts as knowledge only if it bears the right relation to the state of affairs it represents. By studying the causal history of grammars we find that they have been acquired in a way that warrants seeing them as being genuinely representational. A speaker who has the competence to speak fluently stands in the appropriate relation to the structural features of his language to be said to have a representation of those structural features in a grammar. The step from representation to knowledge comes as soon as we say that he knows his grammar: that is, that he bears the right functional (attitudinal) relation to those representations. For, to know something an agent must not only be appropriately related via perception and learning or thinking to the object of knowledge: he must be able to harness that knowledge in the right way.

Now, Chomsky and Fodor maintain that we harness our representations of grammatical relations through recognizing and generating well-formed sentences. They see the same representational state being used in both

---

26 Chomsky, op. cit. (1965), p58.

cases, for it is in virtue of knowing the relevant structural relations in a language that we can construct or deconstruct a sentence.

Where they go wrong is in assuming that if an agent has acquired certain abilities because of certain regularities in the environment of action and these abilities are so conditioned that they can be associated or disassociated in a combinatorial manner, that therefore the basis for those abilities must be unified structures. Why must we suppose that the way knowledge of linguistic relations is harnessed is by invoking an actual physical symbol structure that represents the content of that knowledge? We need a separate argument to establish that the way the agent implements its conceptual knowledge is by using a 'term' in the language of thought that conveys the sense of the concept. Such an account does seem to fall prey to Rylean regress arguments, for in identifying grasping a concept with operating with a formal structure, it attempts to reduce the ability to recognize, say, verb phrases, to the ability to recognize when it is appropriate to use the formal structure. Unless this ability is both simpler and equally robust it suffers the fate of all entity theories of recognition: ad hocness. The level at which we have found robust entities, however, is the level of concepts, of knowledge, of recognitional abilities. Chomsky and Fodor's psychological and linguistic theories would be no less interesting for remaining agnostic about how knowledge is implemented. To go further and make strong claims about the nature of either information or cortical processing simply encourages an attack on the very idea of knowledge level research. For it encourages identifying all arguments against the robustness of the language of

thought as arguments against the reality of knowledge states. By divorcing the knowledge-level from the language of thought, we can salvage what is methodologically valuable in the approach of Chomsky and Fodor, and defer requests for mechanisms until more is known about the nature of information retrieval and processing at the neural level.

Chapter Four

ARE THERE TWO ORDERS IN NATURE?

It is time now to explore in detail what may be the most profound skeptical challenge to knowledge level research. For at least a century philosophers have been arguing that the interpretative sciences -- the sciences which approach man at the knowledge level, as a rational being endowed with beliefs, desires and intentions -- belong to a different category of research than the natural sciences. In the interpretative sciences man is conceived of as a rational, concept manipulating animal whose actions are undertaken knowingly, hence they must be explored with an understanding of the very concepts agents operate with. In the natural sciences man is a parameter optimizing machine, a complex control mechanism. Human behaviour is structural or functional, not intentional; it is something that optimizes certain variables, whether comprehensible to the agent or not, and which proceeds automatically because of internal causal forces. In principle, these approaches need not conflict. And if the knowledge level is to constrain mechanical design they must not. Actions that are intentional may be redescribed as behaviour that optimizes certain functionally defined parameters. Indeed most philosophers today see the differences in the two sciences as purely linguistic. Since an event is explained only under a description, that is, from a particular descriptive perspective, the more ways we have of describing an event, the more accounts we may find to explain it. Thus although most mental events or states are currently described in a vocabulary which

emphasizes their rational, intentional or semantic connection to antecedent and subsequent events, they nevertheless refer, according to materialists, to particular portions of space-time which can be described in physical terms. It follows that as long as every particular action and every particular mental event can be given a physical description, and every physically describable event can in principle be given a physical explanation, rational and mechanical explanations will be compatible. For it is a principle of logic that two statements or explanations cannot both be true of something unless they are compatible.[1] Hence, if rational explanations have a truth value, they must be compatible with mechanical explanations -- they are true of physical events under a different description.[2]

Unfortunately skeptics[3] are not so much concerned with compatibility as with commensurability. How can knowledge constrain design if the two belong to different ordering systems? Explanations are concerned with order. Likewise, theories are concerned with ways of classifying and ordering events into networks of relations. Although there is nothing absurd in a metaphysics which allows one and the same system to be characterized as supporting radically different

---

1 See Ernest Nagel and James R. Newman, Godel's Proof (London: Routledge and Kegan Paul. 1971) for a readable account of the basis of this principle; esp. chap. II.

2 This version of compatibilism presupposes a theory of causation, a theory of explanation and the assumption that all actions have physical realizations. It was Davidson's contribution to have unified these theories in a comprehensive theory of mind. Of particular interest are his articles "The Logical Form of Action Sentences", "Causal Relations", "Mental Events", and "The Material Mind", all reprinted in his Actions and Events op. cit.

3 See, for instance, Alan Gauld and John Shotter, Human Action and its Psychological Investigation (London: Routledge and Kegan Paul. 1977) esp. Part I.

patterns of order, it will pose a problem for knowledge level theorists if the two orders are incommensurable.

Accordingly, the problem I shall address in the remainder of this thesis arises when we challenge the hypothesis that nature is mechanistic all the way up. Problems emerge when we deny that theories at the knowledge level order systems into collections of mechanically interacting robust entities. It is true that in earlier chapters I spoke as if theories at the knowledge level are sufficiently formalizable that we can test a knowledge base for its sufficiency in generating an input/output schedule. Since formalization implies mechanism, it naturally followed I was assuming that theories at the knowledge level were in principle convertible into mechanical theories. I do not believe, though, that this condition holds always. Questions of sufficiency and necessity loom large at the knowledge level because of their tremendous utility: they promise to offer us existence proofs and impossibility proofs of computational mechanisms. If we know that a certain body of knowledge is sufficient, in principle, for a task then we know that mechanisms with the input/output table in question, mechanisms with, say the ability to speak in pig Latin or to interpret blood samples, are in principle constructable. Similarly, if we can show that the knowledge required for a task is <u>unbounded</u> -- that competence in a task can be regularly achieved only if the system can access arbitrarily large numbers of facts, principles or equations -- then the task is virtually impossible; it must be limited before a machine can undertake to perform it. The hope of making such discoveries is the primary incentive for research at the knowledge level. Yet for all their

advantages, knowledge level theories that provide necessary or sufficient conditions of task performance are the exception and not the rule. Most of the tasks that are truly human are open-ended; they call on such vast amounts of information that cognitive scientists, at present, have little idea of how to manage the requisite knowledge. Moreover, the belief that we will one day mechanize and totally specify such knowledge states seems to be an unlikely dream. This is the claim I wish to consider.

Peripheral vs. Central Cognitive Science

A distinction has become current in cognitive science between faculties that have to do with the periphery of the mind, the interface between mind and world, faculties such as vision, audition, manipulation and grammar; and faculties such as reasoning, concept formation, thinking, and problem solving, which have to do with the formation, organization, manipulation and management of attitudes. Peripheral faculties have been the ones most carefully studied to date, and the ones I appealed to in my defence of realism. In part this is because it is possible to define the job of peripheral faculties quite precisely; we often know exactly what states of occurrent knowledge our peripheral faculties should deliver. We also understand many of the logical and empirical constraints, or precedence relations operating in peripheralist task environments. Accordingly, we can explore methods by which our faculties might create those occurrent states. We can identify states of tacit knowledge that would be sufficient for the task, for the task itself is highly bounded: the kinds of events, or factors that can interfere with the faculty's successful performance are restricted, and

so the faculty need not be endowed with unduly large amounts of tacit knowledge.

Central faculties, by contrast, have been less well studied and with less success. As Brian Smith wrote

> ...the scientific investigation of the central aspects of mind is plainly in trouble. It is not as if a once healthy project has come on hard times; early gains now seem less impressive than they once did and it has proved remarkably difficult to make additional progress. Admittedly, there are some areas under active study: non-monotonic logic, ... In general, however, whereas central questions used to generate considerable enthusiasm, they are increasingly being viewed with caution and hesitation. The concerns that Dreyfus raised in 1972, we have slowly come to admit, were more important and more penetrating than was first recognised.[4]

In this chapter and the one which follows it, I shall shift my attention from peripheral faculties to central faculties. The complications this introduces descend from the inescapable fact that the mind -- or, at least, the higher faculties of the mind -- are inexorably linked to society. Our central faculties are so connected with language, consciousness and public norms that the entire program of exploring central cognitive tasks becomes enmeshed with understanding the way society understands tasks. This dependency on social norms and conventions does not itself upset the cognitive program: there is no reason, a priori, why society's norms should not define a class of perfectly mechanistic well-structured tasks -- witness arithmetic. Thus, given that our central faculties are

---

4 "The Seven Percent Solution", unpublished manuscript, p.1.

intricately related to our social conventions, it does not follow from that fact alone that our central faculties are not principled faculties which can be studied at the knowledge level and theorized about in a manner that allows speculating about mechanisms. The difficulty is, rather, that most of the tasks that rely on our central faculties are more ill-defined than arithmetic, and our society has less clear cut norms and principles about their proper performance.

The problem is that, since knowledge is attributed to a system by 'computing' the inverse function of rationality over behaviour before we can ascend to the knowledge level, we must have a clear conception of rationality. So far, I have been assuming that we have this clear conception, and moreover that it is based on a sufficiently small set of principles and norms that it ought to be possible to decide for any possible action A, in a given environment E, whether or not A is a rational action. If, as theorists, we know what counts as rational in E, then we may be able to work backward to find the knowledge the agent would find necessary or sufficient for task proficiency. The circle is complete when, on the basis of our knowledge level research we consider the sort of mechanisms that might satisfy the constraints on processing which our knowledge theories impose.

What makes the field of central cognitive science confusing is that we attribute to people beliefs, desires and knowledge even in domains where we cannot specify the rationality function. In these domains we seem to be dealing with occurrent knowledge or knowledge that could be occurrent. Hence we have a good idea of what we know. Furthermore, our attribution of knowledge to others does not seem to require that the rationality function be well-defined. It is sufficient that we

share a conception of how to act, or change their minds, etc., to be able to attribute each other knowledge. Yet if our capacity to recognize what is humanly reasonable is not built on a small set of principles: the principles of rational belief management or rational choice, then those other capacities may not be based on a small set of principles either. Unlike vision or grammar, there may be no well-ordered and formalizable system of rules regulating the process of belief change or of rational choice. Clearly there must be at least some principles regulating choice and belief change. But these may so underspecify the ability that we are driven to search for different explanations: hermeneutic explanations, Verstehen explanations and the like, which allegedly are based on a different conception of order than that enshrined in mechanistic explanations. The great problem of central cognitive science, then, may be due to the specific nature of human rationality. If it seems that our concept of rationality applies to environments where there is no effective procedure for defining what is and what is not rational then our capacity to act in a 'recognizably' rational manner may not be a principled one. Put rather differently, the order we recognize in action, thought and desire may not always be expressible in the language of mechanism.

To get a general idea of the concepts of order engendered by the rational and mechanistic approaches let us begin by considering the nature of mechanistic and rational explanations.

Mechanistic Explanation

All explanations, whether mechanistic or rational, assume that certain properties or processes constitute a 'bounded' system where events

outside the system can be assumed to be (causally) irrelevant to the temporal flow of events inside. In mechanical explanations these internal events are logically independent and functionally or nomologically related. They constitute a well-defined internal network. Accordingly, if a system is mechanically ordered we assume that it is both sufficiently determinate and sufficiently insulated from outside influences that the temporal pattern of relations between its states can be economically described in terms of mathematical (or qualitative) functions of the form $P = f(Q)$. Systems that are governed by functions in this way are called machines.

The basic premise of the mechanical approach is that all processes in mechanical systems, no matter how complex, can be decomposed into a temporal series of elementary changes over elementary states, such that any change can be explained as nothing more than the effects of a set of quantitive (or formal) laws determining the behaviour of a basic set of parameters. The vision is of a system changing state moment to moment, as determined by a set of well-defined operations or processes. At any instant, the system is in a definite internal state. Given the kind of processes that can occur in the machine, the internal state determines the next internal states and outputs that are accessible. When an input is presented to the system, the null set being a possible input, one of the accessible states becomes actualized.

In effect a mechanism is a production line of sorts: the result of performing one operation becomes part of the overall 'input' to the next, and the system marches through time changing from state to state in strict accordance with the rules governing machine behaviour. See Figure 1.

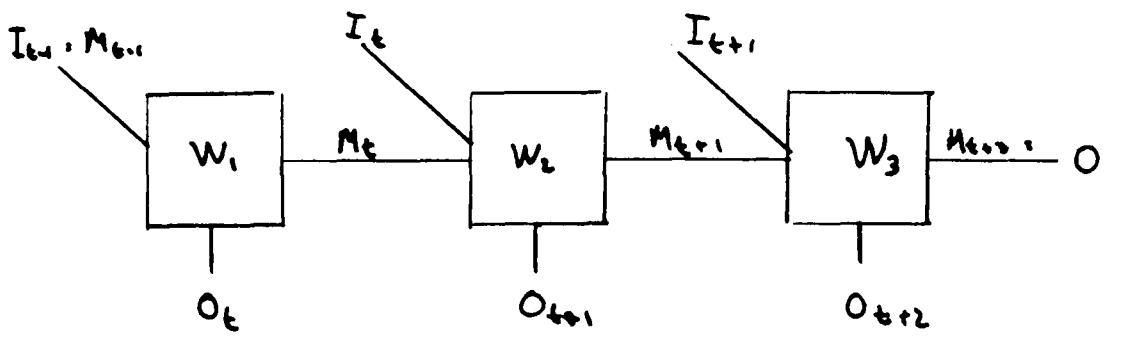


Figure 1

We can formally capture this idea in the language of automata theory.[5] An automaton is a system which:

1) can assume only one of a finite number of internal states at any moment;

2) can produce only one of a finite number of output states or behaviours at any moment;

3) is sensitive to a finite number of input states or signals; and

4) functions according to a finite set of empirical or non-logical laws or rules.

Output, input and internal states are each represented by an ordered n-tuple of logically independent predicates: output by the term $O(o1,o2,...on)_t$ (where t indicates the moment the predicate applies), input by the term $I(i1,i2,...in)_t$, and internal state by the term $M(m1,m2,...mn)_t$. We call the pair $(I,M)_t$ the machine situation at t. The rules which specify the changes that result from each possible machine situation are called the machine rules. They state which output and internal state will arise given input and existing internal state. Represented formally, they show for a given

---

5 For a useful introduction to automata theory see Minsky op.cit., esp. chaps. 2, 4; and Robert Wall, Introduction to Mathematical Linguistics (Englewood Cliffs NJ: Prentice-Hall. 1972) Chap. 10.

$(I,M)_t$ what $(O,M)_{t+1}$ is. When the trajectory of state changes a system follows is simple, and each $(O,M)_{t+1}$ follows from $(I,M)_t$ by the same function f, we can economically express the law governing the system in familiar equation form as $P = f(Q)$.[6] When the system is more complex and the functions governing $(O,M)_{t+1}$ vary with the values of O and M, it is easier to represent the dispositions of the system in a program, or set of condition-action rules, or in a machine table, which states the disposition of the machine for each possible $(I,M)_t$ state. Figure 2 contains an example of a typical machine table. Beside the table the same rules are exhibited as a program and set of condition-action rules. Henceforth I shall assume that any mechanistic process can be represented in one or another of these automata formalisms.

Machine Table

I \ M	2	5	7	4
1	1,2	1,4	1,6	
2	1,0	3,7	5,9	
3	3,2	1,2	1,4	

Condition Action Rules

If (I is even, M odd) → (O is absolute differen
M is sum)

If (I is odd, M even) → (O is I, M remains)

If (I odd, M odd) or (I even, M even) →
(O is I, M is absolute difference)

Figure 2

Until recently it was not customary, in philosophical discussions of mechanism, to dwell on concepts drawn from automata theory. The models of mechanistic systems that philosophers talked about were taken

---

6 More precisely the equation $P_t = f(Q)_{t-e}$. But as e may be made arbitrarily small we can treat the functional relation to be instantaneous.

from physics or chemistry. They were not systems regulated by dozens of distinct laws, as one finds in the biological or psychological sciences; they were governed by a few high level laws, such as Newton's laws of motion, or the classical laws of thermodynamics. So the logical form of accounts explaining the behaviour of these systems was easy to represent in deductive nomological terms. For example, to explain why system a changed state from P to Q, it was thought sufficient to list the relevant laws plus initial conditions. That is,

$$\frac{\begin{array}{l}Pa \\ \forall x(Px \rightarrow Qx)\end{array}}{Qa}$$

If more laws were involved, or initial conditions were more complicated, these were simply added on as extra premises.[7] No real effort was made to find an elegant way of displaying all relevant laws and causally relevant conditions of a mechanical system. Since the systems considered were governed by a few rules it was a fairly simple matter to apply the model to new conditions.

The advantage of using the language of automata theory to represent mechanical processes is that we can readily accomodate mechanisms of arbitrary complexity. When the subject of study is as complex as mind this feature is all important. The great methodological problem of psychology today is to find a level of description to characterize the processes in the brain which is simple enough to allow models to be built that are easy to understand yet complex enough to allow those models to generate detailed predictions of behaviour. Assuming the

---

7 cf. Carl Hempel "Aspects of Scientific Explanation" in his book of the same title (New York: The Free Press. 1965) particularly section 2.3.

brain to be a mechanism of fantastic complexity, we can expect the number of laws governing its state changes, even at high levels of abstraction, to be enormous. Even if psychologists were to eventually offer a unified theory of intelligent behaviour -- a prospect unlikely to be realised -- they would undoubtedly have to represent the neural control system as a system regulated by a staggering number of specific rules. By adopting automata formalisms, psychologists are free to exploit any discoveries made in the field of software engineering -- the one subject where attention is paid to the organization of complex systems of rules. These discoveries may be as straightforward as those concerning new forms of control structures, as when new types of hierarchical, heterarchical or sequential control patterns are discovered, or they may be as esoteric as mathematical theories of program verification and reliability, which may shed light on the type of rules or rule sets which would be most likely to be evolutionarily selected for. There is no denying the value of choosing to phrase theories in a language about which more is being discovered daily. The formalism of the D-N[8] model is simply too impoverished to aid in the sorts of mechanistic studies we need in order to advance in our understanding of complex control systems.

The second reason for preferring the formalisms of automata theory to D-N formalisms for discussions of mechanism is that the notion of an automaton does a better job of capturing the intuitive idea of mechanism than the notion of what is explicable in D-N terms. For

---

8 Henceforth, in place of the phrase deductive-nomological, I shall use 'D-N'.

instance, Hempel's account of explanation[9] recognizes all cases of inductive subsumption as explanations. To explain why this bird is black, we might be told 'Because it is a raven and all ravens are black'. The 'because' here is not a mechanistic 'because': the law 'all ravens are black', though empirical, is not a quantitative or qualitative function of the sort we are interested in. To circumvent this problem, philosophers historically have taken 'mechanical' to mean 'physical'. Since physics is regarded as the metaphysical basis for determinism and lawfulness in nature, the question of whether a system was described as a mechanical system, reduced to the question of whether that description could be interpreted in the language of physics. If it were possible, in principle, to give a fully adequate account of the system in terms of the concepts and laws available in the physical sciences -- ideally to translate the story told in non-physical language into a story told in the language of physics -- then one would know that the system was mechanical.

But as modern functionalists[10] have pointed out the concept of mechanism is independent of the concept of substance. Even if minds were made of immaterial mind-stuff, mental systems might change state in strict accordance with machine tables. Therefore, although the thesis of the mechanistic <u>materialist</u> might be false -- that is, although it may be false that man is nothing but a complex physical mechanism, differing in degree of complexity but not in kind from other physical mechanisms -- nonetheless, man would not thereby be shown to have a non-mechanical aspect. All that would be shown would be that a

---

9 Hempel op.cit.

10 Most notably, Putnam in (1960) op.cit., and his other papers on functionalism.

science of mental mechanics would require concepts and laws that were irreducible to physical concepts and laws. Consequently by adopting the language of automata theory to represent mechanism, we can capture what is essential to mechanism without presupposing reducibility to physics.

Once we think of a mechanical system as a bounded process which can be analysed into a set of basic elements obeying determinate transformation laws, we can treat as mechanistic any system whose trajectory of state changes are a function of time. All that is required to be mechanistic, then, is that:

A. Each successive state in the temporal trajectory of the system must be a determinate function of last state (including new input), hence individual events must be determined uniquely and precisely by last machine situation, implying that there can be no action at a temporal distance.

B. The temporal trajectory of the system cannot be affected by processes outside the spatial boundaries of the system: all changes are caused by local interaction between parts or between the system and signals (forces) which impinge on part of the system; hence there can be no action at a distance.

C. The set of possible states of the system is recursively enumerable by applying all the transformation rules to an arbitrary state plus allowing the values of input to vary over the system's entire input sensitivity. Thus we can precisely define any internal state.

It is easy to see that the roots of mechanism lie in atomism.[11] Condition C, for example, requires that each internal state be a function of a 'few' elementary states. These elementary states cannot change their dispositions moment to moment; they are atomic, fixed in dispositional structure and relatively permanent states of the system. A stable combinatorial basis is necessary for recursion. Moreover conditions A and B require that atomic elements interact locally. Causation emanates outward; it proceeds by contact.

Historically the conceptual enemy of atomism has always been holism. And in a mechanistic psychology, no less than in a mechanistic physics, holism creates a problem. If a system can be interpreted only as a whole, the best explanations one may find may be global. Interactions may have to be treated as a web which has no preferred partitioning. Moreover, the web may be best construed as continuous, composed of entities varying infinitesimally in size and position. Consequently it might actually be false to say that the system was made up of a finite number of parts interacting locally, because the concepts of local interaction may be undefined. Between every two parts there is a third.

Such fundamental concerns become relevant in cognitive science precisely because the web of belief and desire may be unbounded. Beliefs may 'link' up with other beliefs without end: the whole system

---

11 For an interesting discussion of mechanistic order in physics, and the conceptual link between atomism and mechanism, see David Bohm, Wholeness and the Implicate Order (London: Routledge and Kegan Paul, 1981) esp. pp.172-79.

constituting a web with no definite beginning or stopping point.[12] Cognitive scientists, aware of this problem, have attempted to circumscribe this holism by organizing belief systems into frames.[13] But frames too have proved unsatisfactory unless equipped with pointers which reach outside themselves to additional frames arbitrarily distant in time, space and thought. Hence frames are not atoms either for they are open systems; their dispositions are partly undefined, dependent on the entities filling their pointers.

Holism is the spectre haunting philosophically acute cognitive scientists. Nonetheless, the mechanistic framework defines the horizon of most work in cognitive science. To get a rough idea of the way higher mental processes are in practice characterized mechanistically, consider this account by Hunter of his research on rapid mental calculations by the distinguished mathematician, Prof. A.C. Aitken:

> During 1961, I had the rare privilege of studying a man with exceptional ability in rapid mental calculation... His unusual powers may be illustrated by two examples. He is asked to express as a decimal the

---

12 Cf. Davidson in "Mental Events" op.cit., "Beliefs and desires issue in behaviour only as modified and mediated by further beliefs, desires, attitudes and attendings, without limit." p.217.

13 'Frame' is the term used by Minsky in "A Framework for Representing Knowledge" appearing in P.H. Winston (Ed.) The Psychology of Computer Vision (New York: McGraw-Hill, 1975) and in "Frame-system theory" reprinted in Johnson-Laird op.cit., to refer to clusters of properties that can, with greater or lesser confidence, be associated with a given situation. To be useful, frames must be organized into linked systems that make it possible to co-ordinate information gathered from different viewpoints, so that an agent can keep track of what has changed and what has remained constant in his world, so as to be able to make as many reliable inferences as possible in any given situation. See also Patrick Hayes "The Frame Problem and Related Problems in Artificial Intelligence", reprinted in Bonnie Lynn Weber and Nils J. Nilsson (Eds.) Readings in Artificial Intelligence (Palo Alto: Tioga, 1981); and Margaret Boden, Artificial Intelligence and Natural Man (Hassocks, Sussex: Harvester Press, 1977) pp.305-314.

> fraction 4/47. He is silent for four seconds, then begins to speak the answer at a nearly uniform rate of one digit every three-quarters of a second. "Point 0851063829787234042553191 4, that's about as far as I can carry it." The total time between the presentation of the problem and this moment is twenty-four seconds. He then discussed the problem for one minute and then continues the answer at the same rate as before. "Yes, 191489, I can get that." He pauses for five seconds. "3617021276595744 58, now that is the repeating point. It starts again at 085. So if that is forty-six places, I am right." The second example...
>
> Professor Aitken solves any given numerical problem in a sequence of steps. First he examines the problem and decides the plan or method by which he will calculate the answer: in doing this he typically recasts the problem into a form he can more easily handle. Then he implements his chosen method and, step by step, generates the answer. This step-sequence is evident in...[14]

Hunter describes Aitken's extraordinary skill as the result of having some master plan which is able to decompose a calculation problem into an organized collection of simpler problems which can be tackled sequentially. As Fig. 4 suggests, the picture offered is of a mind equipped with a vast storehouse or library of methods. Part of the mind is triggered to search this library for a method that will likely obtain the solution in the shortest time and with the least difficulty. This is a complex search, for it involves first transforming the problem, in accordance with the rules of each method, into a new format or representation that is easier to handle than that implicit in the original formulation, and then testing if it is the best format. This process is mechanized. After this selection of general approach has been made, the mind or 'control' centre then selects a more specific method to handle the newly recast problem. Again, the new method involves moving step by step through

---

14 "Mental Calculation" in Johnson-Laird op.cit. p.35.

computations. Finally the answer is produced and the control recognizes that a solution has been found and generates output that counts as the answer. See Fig. 4.

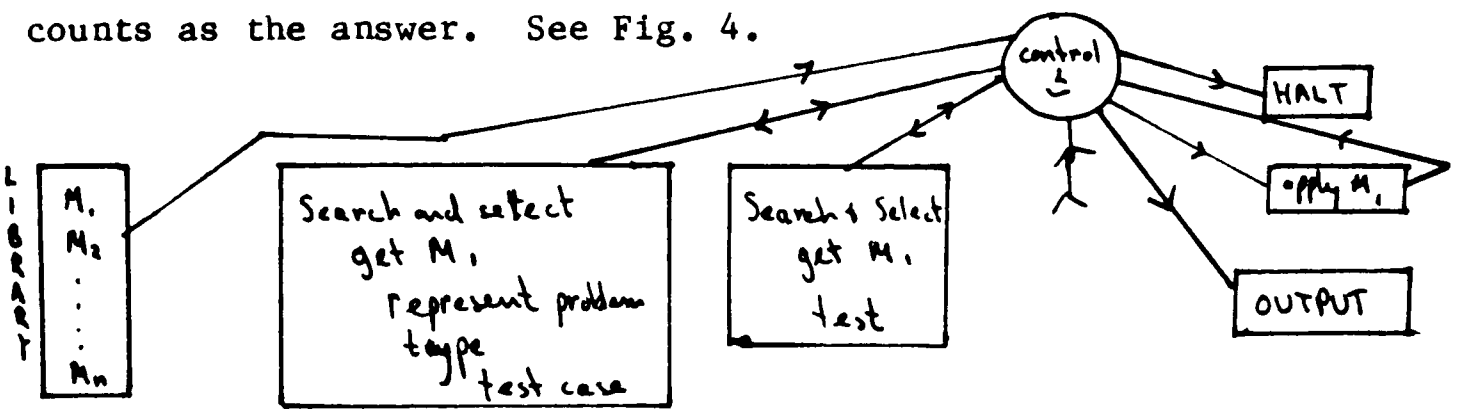


Figure 4

Clearly the mind of Aitken is being represented in the formalism of automata theory. He is described as having a master plan in mind which breaks the overall calculation job into a number of specific smaller jobs, each of which is performed in accordance with definite sub-plans and which carries him along a distinct problem-solving trajectory that satisfies conditions A and B. Had the problem been different the trajectory would have been different. But from the start, the set of all trajectories Aitken could possibly follow was determined by his master plan. Hence condition C was fulfilled as well. The set of his possible mental states was recursively enumerable.

Thus Aitken was treated exactly as if he was a computer. His action was represented as being no different in principle from computer output and his inner thoughts were not elaborated by appeal to his beliefs, etc., but by appeal to the world of formal computation. His arithmetic skill was treated as a peripheral process in which conventions of number theory were internalized. Hence, the explanatory schemas sufficient to explain the behaviour of computers seemed sufficient to explain his behaviour.

Rational Explanation

By contrast, explanations from the rational or hermeneutic perspective -- explanations of central cognitive processes -- rely on a rather different language than the formalism of automata theory. To the extent that there is a single well-defined formalism for rational explanation at all (a claim I shall dispute), it is one which represents the system in question as an interacting set of goals, desires, purposes, beliefs and action states. The principles of interaction are the basic rules of rationality.

For instance, every intelligent being acquires information about its environment and self, and forms beliefs. These sets of beliefs are continually being disturbed by new observationally (or otherwise) acquired beliefs. As new beliefs -- call these prima facie beliefs -- enter the system they must be integrated into the existing belief system, and so interactions take place until a new rational equilibrium has been achieved. Beliefs may be assumed to interact rationally with other beliefs to minimize inconsistency, incoherence, to maximize explanatory potential, and so forth. The rules that govern the attainment of these new equilibria are the rules of rational belief.[15]

Similarly for desire. Intelligent beings must have desire systems. They must be sensitive to states or objects which attract them, and states or objects which repel them. Otherwise behaviour would be pointless, directionless. According to the canons of rational explanation, desires interact with other desires and with beliefs too,

---

15 For an account that assumes the process of achieving rational equilibrium to be formalizable in principle, see Brian Ellis, Rational Belief Systems (Oxford: Basil Blackwell, 1979) esp. Chapter 1.

to create rational chains or hierarchies of intentions or goal structures that drive the agent to act. These goal structures later interact with occurrent beliefs to produce action that is sensitive to the agent's current situation. Desire systems, then,can also be in greater or lesser degrees of rational equilibrium. If too many desires are in conflict, relative to the world the agent believes himself to be in, he may have insufficient direction in life; some desires must be dropped or altered or weakened or suppressed. Accordingly, an ideally rational desire system is one which is in equilibrium under the most acute pressures of internal criticism, concerned among other things with rational attainability, sensibleness, wisdom, aptness and so on. These criteria are all metrics of rational desire.

Rationality also enters our understanding of intelligent systems through our analysis of actions. Intelligent beings are guided by reason. Behaviour is supposed to be recognizably 'appropriate', 'reasonable', 'sensible' or 'meaningful' in the light of the belief and desire systems of the behaving agents. It may be impossible to specify an effective procedure for evaluating the meaningfulness or reasonableness etc. of an action vis-à-vis a belief and desire system, but it is the premise of the rational approach that such evaluations can be made, at least by kindred agents. Thus rational agents may be assumed to have a special competence: they can recognize what is rationally implied by sets of (reasonably well-ordered) beliefs and desires. In their own case this recognition ability is coupled with a (conditional) disposition to act on the consequence of their recognition.

Now if we follow the hermeneutic tradition, the way we ought to explain changes in belief systems, desire systems and action systems is

not the way we explain changes in automata.[16] Rational systems cannot be represented as automated belief, desire or action systems because they fail to satisfy some of the conditions of automata. For instance, in automata, all state changes are the outcome of applying a finite set of rules on internal states plus input. These must apply whether the internal changes are called learning or, forgetting, or whether they are the result of short term thinking. Is it plausible that when a child learns its mother tongue the trajectory of belief states it passes through can be described as the result of applying a small set of rules to antecedent states? Does it not acquire new concepts that lie beyond anything that could be constructed out of its old ones? Following this line, is it plausible that all rational systems contain only those internal states that are recursively enumerable from a finite lexicon of primitive elements? All theorists who believe that conceptual change is more than a rule-governed manipulation of a combinatorial lexicon of 'germ' ideas believe that it is, at least, possible to have creative flashes which elude description as constructions from previous ideas. Mental states may be constructed, but are they in principle recursively enumerable?[17]

---

16 This was the explicit thesis of Gould and Shotter, op.cit.

17 In this respect Geoffrey Sampson in Making Sense (Oxford: OUP, 1980); Charles Taylor in "Interpretation and the Sciences of Man" in Revue of Metaphysics 25, pp.3-51; and Paul Churchland in Scientific Realism and the Plasticity of Mind (Cambridge: CUP, 1979) secs 18-21; have all argued that the set of possible mental states of humans is not recursively enumerable because there are no well-defined limitations on the thoughts and novel ideas a person may have. Thus Sampson states: "It is in the nature of human minds to be always coming up with novel ideas -- ideas which are in no sense implied by anything that has gone before, so that the workings of a man's mind are not predictable as machines or natural phenomena are." ibid. p.1.

If these suggestions are true -- and I shall consider them in depth shortly -- then rational explanations cannot be translated into the language of determinate serial automata. We shall have to regard rational explanations to be a different sort of explanatory schema: a non-mechanistic schema. Belief systems, desire systems and rational action systems will not be representable as mechanistic systems. Of course, this is not to say that the physical systems which realize or embody belief and desire systems, are not mechanistic systems. We have already accepted that the relation between the physical events in these systems are all formalizable in automata language. It is rather the relation between the psychological events in these systems -- the relation between beliefs, desires and actions -- that are being called into question. If they are inappropriately related to be formalized, yet they are necessary to understand the changes in mind endowed systems -- that is, if mechanical explanations cannot supplant rational explanations -- we will be forced to admit the reality of a second order in nature: rational order.

The Rational Order

So far I have assumed that being related in non-mechanistic ways is identical with being related in non-functional ways. A more conventional answer would hold mechanistic order to be what is in principle orderable in mathematical terms, and treat non-mechanistic as what is in principle not so orderable. Proof of the reality of a distinct non-mechanistic order, then, would involve showing that rational processes cannot be interpreted as purely computational (ie. serial automata) processes because they quantify over the wrong kind of

entities to enter into computational-type relations. If changes in, for example, belief systems fail to proceed step-wise according to computational laws it may be because beliefs are by nature unfit to serve as elements in a machine.

To see what would be involved in acknowledging an order in nature composed of intrinsically non-mechanistic elements, let us imagine a world where thoughts, beliefs, desires and actions 'interact' not because of any algebraic or geometrical structure they have -- e.g. having sentence-like shape -- but because they have basic 'inexplicable' dispositions to interact in certain sorts of ways. In such a world the interactions of, say, one thought with another is not to be explained by supposing each thought to have a determinate structure, describable in principle as an n-tuple of properties -- e.g. a string of characters -- which then combines or reacts with other thoughts according to a small number of rules. Rather thoughts would be akin to structureless propositions. When entertained in the right medium these structureless propositions react according to their (causal) natures, but not in any simple rule-like way. Thoughts may react with other thoughts or beliefs, etc., in such diverse ways, that we simply cannot find a small set of rules that governs their dispositions.

The picture of the mind we are envisaging, then, is of a domain where interactions between entities are not systematizable in a mathematical formalism. Unlike Peirce, we are not viewing the rational domain as a space where thought follows thought according to one or

another formal rule of inference.[18] Such a space is eminently mathematical, for each thought in Peirce's system is isomorphic to a formal string of characters. In the world of thoughts we are considering, interactions may be more idiosyncratic, depending on the specific (causal) peculiarities of the thought involved.

Although I shall have more to say about the non-formalizable nature of thoughts shortly, I feel obliged to emphasise that the view I am offering is not anti-physicalistic in any interesting sense. It is assumed that each thought token is embodied in a complex mass of physical processes. These processes interact in accordance with physical law. So naturally these physical processes can be formalized. We presupposed that fact in granting that all physical theories can be translated into automata language. But although for each physical process $p_1,\ldots,p_n$ in system S capable of implementing the thought that P and each physical process $q_1,\ldots,q_n$ capable of implementing the thought that Q there may be a machine rule (i.e. a physical law) specifying that in S, a given p_i leads to a given q_j (when in fact P does imply Q), nonetheless it is unlikely that there could be a single machine rule (physical law) covering all the possible ways of implementing the physical processes corresponding to the mental process

---

18 Peirce's anticipation of modern cognitive science is not generally appreciated. Yet it was among his central theories that "we must, as far as we can.... reduce all kinds of mental action to one general type.... This is no other than the process of valid inference... we can, without any other supposition than that the mind reasons, reduce all mental action to the formula of valid reasoning... something, therefore takes place within the organism which is equivalent to the syllogistic process." "Some Consequence of Four Incapacities" reprinted in J. Buchler (ed.) Philosophical Writings of Peirce, (New York: Dover, 1955), pp.230-31.

of inferring Q from P. This point, emphasised by Fodor and others[19] is that what corresponds at the physical level to what we call, at the psychological level, 'inferring Q from P' may be a heterogeneous and unsystematic disjunction of predicates in physical theory. Thus although there may be a physical event corresponding to a particular occasion of 'thinking that P' which leads to a physical event corresponding to 'thinking that Q' in strict accordance with some physical law (plus boundary and initial conditions), there may nonetheless be an indefinite number of diferent physical pathways by which this could come about on different occasions. Thus there need be no physical property which all these different pathways share in common. Each path may be an individual case. Hence, although physical theory applies to all instances of 'inferring Q from P', it does not provide a theory about these instances. If there are regularities to be found between inferences they will have to be found at the psychological level. There is no way of proving that such regularities must be formalizable by appeal to the formalizability of physical theory.

I think the reason cognitive scientists assume that the domain of thoughts and beliefs, etc., can be formalized is that they suppose that it is possible to discover a formal theory of language use capable of explaining all our intuitions about rational linguistic connections. Since the only known way to study thought is to describe it in terms of the structures and properties of some language or other, it is usually assumed that if we can discover a formal theory of the

---

19 Jerry Fodor (1975) op.cit. pp.20-6, and Jaegwan Kim "Supervenience and nomological irreducibility", in American Philosophical Quarterly, 1979.

ways sentences can be rationally associated, ie, deductively, inductively, 'abductively', analytically, praxeologically, and so on, we will ipso facto have a formal theory of the ways thoughts can be reasonably associated. Most cognitive scientists, on being pressed, will admit to a Lockean vision of thought in which:

(1) words or word groups are associated with ideas in the mind;

(2) ideas are held to combine in the mind in ways which resemble the ways in which words or word groups are combined in sentences; and

(3) thoughts are combinations of ideas, (or a specific attitude to such combinations);

Now in last chapter I argued that concepts are dispositions and abilities, not symbols in the head. Thoughts and attitudes, therefore, are the result of exercising conceptual abilities. Accordingly, a thought need not be structured in the sense of being composed of several distinct elements, ideas -- whose categorical basis is somehow isomorphic with their causal potential. Nowhere in my brain must certain symbols be manipulated for me to have the occurrent thought that Bouillabaisse is overrated. It is sufficient that certain conceptual abilities be activated in the appropriate manner and order.

To be sure, thoughts must have at least as much semantic structure as sentences in the sense that they must have the requisite dispositions to enter into sentence-like relations with other thoughts. There is a non-contingent parallelism between the relations holding between the that-clauses which give the contents of thoughts and those thoughts themselves. The question is: Can those relations be formalized? What grounds do cognitive scientists have for thinking that our intuitions of valid deductive, inductive, abductive and praxeological inference can be formalized and hence mechanized? This

is the question with which I shall be concerned in the remainder of this thesis. In the next section I consider the prospects for mechanizing rational belief management in accordance with formal rules of inference. In the subsequent chapter I consider the chances for mechanizing practical reasoning.

Rational Management of Beliefs

The rational management of belief ranks among the critical functions which our reasoning system must perform. Intelligence depends on rapid access to reliable information. As observational, testimonial and theoretical information flows into our reasoning system it must be 'rationally integrated' with previously held beliefs: the new information must interact with established beliefs to create a rational equilibrium which maximises such ideals as consistency, coherence and explanatory power.

Now it is a tenet of the hermeneutic approach that the trajectory of changes our belief system undergoes can be recognized as rational. Whereas the mechanist believes there is a set of formal rules governing these transitions and that these are well-defined over linguistic entities, the rational man theorist remains skeptical about the chances of formalizing the dynamics of belief change.

A belief system is a set of beliefs. So far I have not elaborated what I take a belief to be. Some have argued it is an attitude to a propositon grasped;[20] others suggest it is a disposition to use

---

20 E.g. Gottlob Frege in "The thought: a logical inquiry", reprinted in P.F. Strawson (ed.) Philosophical Logic, (London: Oxford Univ.

public sentences in certain ways;[21] still others maintain it is a relation the believer bears to a set of possible worlds.[22] Whatever one's theory of the nature of belief it remains uncontested that to describe a belief system we must say which sentences of a language the agent accepts as true, which he accepts as false and which he remains undecided about. The support for this intuition can be found in the principle of expressability, which asserts that there are no ineffable thoughts. According to adherents of this principle, whatever can be thought or believed can be stated -- not necessarily in languages as they are now, but in principle. Any language may be too impoverished in word or structure to currently capture all beliefs. But this is a contingent fact. In principle any natural language can be sufficiently enriched by introducing new terms or other devices to express all thoughts. For if one could not publicly represent one's thoughts there would be certain thoughts that were necessarily private. Yet how would the concepts involved in these thoughts be learned? If it be replied that they might be innate, or learned privately, then we need to be told how the agent makes sense of the thoughts to itself. This becomes paradoxical if we confine our attention to thoughts about the world, where possession of a concept is at least manifestable in recognitional

---

Press, 1967) pp.17-38; and his spiritual heir Alonzo Church in "On Carnap's analysis of statements of assertion and belief" Analysis, 10(1950), pp.97-99.

21 This I take to be Quine's position in Word and Object, (Cambridge: MIT Press, 1960).

22 E.g. Robert Stalnaker. "Propositions" in Alfred Mackay and Daniel Merrill (eds.), Issues in the Philosophy of Language,(New Haven: Yale Press, 1976).

capacities directed on items in the world at large. Thus, if we consider belief systems about the external world we may assume that corresponding to every belief there is a sentence in some idealized language which expresses the content of the belief and to which the agent is disposed to assent with greater or lesser vehemence.

Given the principle of expressability it follows that all the 'logical' relations a belief or other attitude can enter into -- all 'rational' relations of a deductive, analytic, inductive or praxeological nature -- will be mirrored in the 'logical' relations a corresponding sentential expression can enter into. In fact, all such relations are defined first over sentential expressions and then parasitically over beliefs. Thus, for example, if I believe that p, and I believe that q, and as a matter of logic, p entails q, or q entails p, then my two beliefs can in principle be deductively related, because there is a formal rule mapping over their sentential expression which sanctions the inference. Conversely, if there is no deductive logic (operating on sentences) which shows them to be formally related, they cannot be deductively related. Again, if I have in my belief set p, p', and q, and p, p' are related to q as evidence is to hypothesis, then my beliefs may in principle be inductively related, otherwise not. Without the sentential expression of belief, inductive relations would be undefined. Again, if I believe p, I desire q and intend a, and these are praxeologically related as a reason is to its correlative intention to act, then my relevant belief and desire may in principle be rationally linked to my intention. The proof is that we may construct a practical syllogism expressing in sentence form the belief, desire and intention relation.

We may be sure that there is a complex interaction between what we believe we are justified in inferring on the basis of having learned to infer sentences from other sentences, and what we are disposed to infer unconsciously when managing our own belief systems. Belief systems are meant to be in rational equilibrium. They are meant to provide us with the best theory of the world, and so we wish to subject them to the greatest possible internal criticism and evaluation. Naturally the principles we have learned publicly help in these matters. But beliefs, after all, are more than sentences. They literally engage the perceiving, reasoning and learning systems, and have causal powers which utterances have only in the most parasitic and bloodless fashion.

Because of the psychological pressures beliefs may come under in virtue of participating in a living, changing cognitive economy, changes in belief systems may not always be due to the logical pressures which we would admit our beliefs ought to be subject to were we to display them to the cold light of public scrutiny. As Haugeland has argued, thoughts and beliefs

> ... are embodied in a 'system' that provides normal channels for them to interact with the world, and such that these normal interactions tend to maximize the "fit" between them and the world; that is, via perception, beliefs tend towards the truth; and via action, the world tends towards what is desired. And there are channels of interaction among thoughts (various kinds of inference) via which the set of them tends to become more coherent, and to contain more consequences of its members.[23]

Evidently, beliefs systems are part of the adaptive machinery which evolution has equipped us with in order to enhance our 'fit' with a

---

23 John Haugeland "Programs, causal powers, and intentionality" in Behavioural and Brain Sciences, (1980), 3, pp.432-3.

rapidly changing world. Yet they do not stand alone: it is in conjunction with other components of our cognitive system that they contribute to adaptiveness. This raises a problem. How can we be certain that a system with limited computational resources is best off with a belief system that aims for instance at maximum consistency? In evolution, clever trade-offs often determine the success of a species. Perhaps increased explanatory power in belief systems is more adaptive than increased consistency in human environments. In fact, why should belief systems strive to attain norms which in the public world we regard as desirable? Can we be confident that these norms are rational ideals?

What makes this problem difficult is that, among other things, we cannot be certain of the precise role beliefs play in intelligent systems. We tend to assume that they may play the same role as that which 'accepted' sentences do in public institutions. But sentences, unlike beliefs, link up to the world only through humans who use them. Their intentionality is derivative. They have few, if any, of the causal powers of genuine beliefs and thoughts.

Despite this prima facie difference between beliefs and sentences, rational inference is defined first over sentences then over beliefs. The question we must address then is this: what prospect is there of finding a single formalism that is powerful enough to support the various systems of inference rules necessary to accommodate all our intuitions of rational inference? Is there a single language that can formalize our intuitions of deductive, inductive, analytic and practical inference? A single language is necessary because we want our account to provide the basis for an explanation of how a single

faculty operates. It won't do to say that the belief management faculty operates as if it represents its beliefs in L_1 when it alters the belief system through deduction, L_2 when it relies on inductive rules, L_3 when it relies on semantic rules and so on. The point of the faculty is to establish rational equilibria under critical scrutiny. If there is no unique language over which we can define what 'critical scrutiny' means, then it is undefined. It would be as if we defined addition over numbers, paraphrase over sentences, and themes over pictures, then without showing number, sentence and picture notations to be equivalent, we defined a new relation which blends all three relations without proving there is a single notation over which it can be defined. What would such a relation mean? How would we use it?

One reason for doubting that such a lingua universalis can be discovered for belief is that it is not at all clear that each of our intuitions about deduction, induction, explanatory plausibility and so forth can be satisfied even in their own tailor-made notations. Two facts point to this conclusion. First and most importantly, there is no single logic of deduction, induction or explanation to appeal to. Take deduction: one can, if one wishes, talk of a canonical rendition of the laws of logic. But the logician who regards his task as that of writing down all the general rules implicit in our logical appraisals is more likely to suppose that there is more than just the canonical way of being inconsistent or of making valid and invalid steps in an argument. There is nothing to prevent someone from stipulating that a class of 'informal' deductions shall henceforth be called exemplary and construct his formal theory to accommodate those. He might, for

instance, state that his concern is with types of inconsistency, validity and invalidity that are topic-neutral, not confined to particular kinds of subjects.[24] But there is bound to be an element of arbitrariness and vagueness in any criterion adopted.[25] Consequently, formalisms abound, and the decision as to which logic to accept as the deductive logic can only be settled by showing that one does better in satisfying pre-formal intuitions. Each formalism serves as a theory about human deductive behaviour. Competing theories therefore are adjudicated on grounds of empirical adequacy: hardly a reason to treat one as the canonical theory, the norm-setter toward which all deductive inference should aspire.

The second ground for suspecting that deduction may not be univocally formalizable is that in matters of belief, domain-specific rules of reasoning are as important as content-independent rules. Since many of our arguments in natural language rely on our knowledge of the meaning or reference of our terms, we often recognize as valid, arguments which can be formalized -- if at all -- only in formalisms which assign semantic structure to predicates. The predicates of English, for instance, are sufficiently rich in internal structure to allow us to infer from 'John is a bachelor' that 'John is featherless' or that 'John is older than 10'. Such 'logical' relations between sentences have their image in relations between attitudes. Clearly what is required is a universal notation that captures not only the

---

24 The term 'topic-neutral' is Ryle's; "Formal and Informal logic" in Dilemmas (Cambridge: Cambridge Univ. Press 1954).

25 See Susan Haack Philosophy of Logic (Cambridge: Cambridge Univ. Press, 1978) sec. 1-2, for a lucid discussion of the vagueness of 'Logic'.

semantic structure of our concepts but their dispositions to participate in inductive and other inferences. But what chance is there for that if we have competing notations for inferences as commonplace as deduction?

To make matters worse, belief management does not involve belief expansion alone; beliefs sets must also be pruned: they must be kept fit with respect to the changing information made available to the agent. That means the process of updating belief sets is non-monotonic; we discard old conclusions in favour of new evidence. In particular, because our belief set aims at achieving explanatory plausibility, we find we have to reject beliefs periodically and to re-orient others. The trajectory of changes this creates has peculiarities that have eluded description at the sentential level.

For example, Quine,[26] has stressed that we can reject any of our beliefs at the expense of making suitable changes in our other beliefs. When an observation shows that a system of beliefs must be overhauled, it leaves us free to choose which of those interlocking beliefs to revise. Beliefs face the tribunal of observation not singly, but in a body. Even observational beliefs, the customary basis for our higher level beliefs, are not immune to revision: observations that are uncongenial to a system may be dismissed as error or hallucination. Hence, there are an indefinite number of equivalent revisions of our belief system that may be equally rational.

---

26 Most notably in "Two Dogmas of Empiricism" reprinted in From a Logical Point of View, (Cambridge: Harvard Univ. Press, 1953).

Philosophers such as Harman,[27] have tried to show that underpinning these changes there is a general inferential mechanism: inference to the best explanation. Assuming our belief system at any time to be in rational equilibrium, we revise it, upon receiving new input, by changing it to maintain something we may call explanatory plausibility. We may think of our belief system as an intricate homeostatic device, dedicated to maintaining the system at a steady state of explanatory plausibility. Disruptions from the steady state cause compensatory reactions and the system returns, in a few oscillations, to a position of minimal stress. As Peirce would say, the irritation of doubt and disbelief leads to inquiry whose aim is the cessation of irritation. Harman's contribution was to study the way people do revise beliefs in an effort to extract an inferential mechanism guiding belief revision. And yet inference to the best explanation remains ill-defined. Quine and Ullian[28] note five virtues which count toward plausibility and which a hypothesis may enjoy in varying degrees. But nowhere do they suggest that these five, conservatism, generality, simplicity, refutability and modesty, which not uncommonly are in 'tension', constitute the basis for an effective procedure for belief revision. On the contrary, they vary in their importance from situation to situation, person to person. Revisions have costs and the degree to which we bias one as opposed to another possible revision is a function, among other things, of how dearly we

---

27 Harman's views are summarized in his book Thought (Princeton: Princeton Univ. Press, 1973). Of particular relevance to us are Chaps. 3 and 8.

28 In The Web of Belief (New York: Random House, 1970) Chap.5.

cherish the previous beliefs that will have to be sacrificed. The list of costs may be extended. Since not all believers, however, assign equal importance to each costing factor; and since there is no unambiguous standpoint from which to judge the relative importance of these costing criteria, there is no hope for an effective procedure. Only wishful thinking could drive a reasonable person to believe there is a universal algorithm for belief. We can say that all believers have a sense of 'overall plausibility'; we can even say they rely on this sense as a regulative ideal in their activity of managing belief. But there are simply no grounds for supposing that all believers share certain basic intuitions about what it is to be rational in belief management, or that these intuitions are sufficiently precise and consistent to lend themselves to formalization. Notoriously, qualified scientists disagree over the 'plausibility' of competing theories. And notoriously formalisms abound in inductive logic, none gaining universal adherence from the informed community. Yet if there is lack of consensus at the social level, why suppose there is a universal notion of rational inference underpinning belief management at the personal level?

Such considerations, drawn from a study of the behaviour of allegedly rational agents in their social activity, point to a general uncertainty about the exact nature of rational inference. Because agreement is often reached, there are areas of thought and discourse where we have the illusion that we share a common body of rules defining rationality. But disagreement is also common. And in specific attempts by philosophers to formalize the rules underpinning our practices, the one thing which stands out is their conspicuous failure to elicit consensus from the informed community. It is not an

accident that there are so many formalisms for logic: the phenomena studied form a heterogeneous class, driven by heterogeneous intuitions.

A Wittgensteinian Problem for Belief Management

I want to pursue a second point of argument now as further support for the skeptical prognosis that a science of our central cognitive faculties can expect little precise direction from research at the knowledge level. It should be emphasised, though, that to support the skeptic is not to argue that knowledge level research is pointless. Clearly, the more we learn about the ideals of the management of belief, the better our position to speculate on particular psychological mechanisms which might implement those ideals. The problem is that as a group we have confused ideas about these matters, and we are confused where as a group we can expect no better from individuals. Unlike our peripheral capacities, where knowledge is 'wired into' the structure of the faculty, our central faculties are more re-programmable, they keep pulse with shifting norms and conventions. Further, the greater the diversity of tasks each central faculty must be able to participate in, the less likely we are to discover much of structural interest by studying any one task environment. To learn about our central faculties we have to ascend to a higher level of abstractness, evaluating what knowledge or properties of knowledge are common to all the different tasks a given faculty might perform.

The Wittgensteinian objection to central cognitive science which I shall discuss plays on the vagaries of consensus endemic to rational inference. Inference in the public world is a matter of obeying a practice, tacitly acknowledging a social norm. If we can show that

there are no such things as deterministic rules regulating inference -- that conformity to a practice does not require knowledge of rules so much as an ability or disposition to bring one's conduct into conformity with the conduct of others -- then the very idea that we might ever succeed in regimenting our intuitions of rational inference would be misguided, for our intuitions may shift without our knowing it. If the ability to infer in a manner generally recognized as rational is a basic ability to do what other people similar to us do, then as long as we all march in step, our steps may change if viewed externally, but seem constant to us.

Criticism at this level of abstraction constitutes a wholesale attack on mechanizing inference. In principle, it applies as much to the rules of arithmetic as to the rules of belief management. But whereas if arithmetic practices deviated considerably, our buildings and bridges would collapse, it is less clear what would happen if our belief management practices deviated. The value, it seems to me, of considering Wittgenstein's argument is that it reminds us that the roots of conformity lie in an indefinite background of conditioning, learning and biology. It is possible, therefore, that machine simulations of mental processes may teach us little about our minds because they have the wrong architecture.

Wittgenstein argued that it is sometimes otiose to explain why a person makes a certain judgement, say, that a is F, or that G follows from H, by claiming that he has a rule which tells him how to make such judgements, because there may be no way of deciding whether or not the rule he is allegedly using at present is the <u>same</u> rule as the one he

allegedly was using in the past.[29] If, however, it is impossible to decide whether an agent is using an old rule or a new rule, there is no point in saying there is a definite rule that he is following. Putting in rules where really there are none gives the world an air of determinacy which it does not have.

Consider how we ordinarily think of rules. On the standard view, to be able to appropriately apply a rule an agent must know the sorts of conditions in which it is correctly applied. Each rule has its conditions of application which define its use. Change its conditions of use and the rule changes. Now suppose the agent finds himself in conditions where he cannot readily decide whether he should apply the rule. Suppose also that he makes a snap decision; he decides on the spot what to do. What, on the standard view, shall we say? Has he changed the rule? We can expect one of two replies. The first might run like this. "Yes he modified the rule to accommodate the new case. It needed to be broadened. His conception of what constitutes appropriate conditions was simply too narrow. So he changed the rule". The second is rather different. "No he didn't change his conception of the rule or the conditions under which it is appropriate to apply it. It is rather that he didn't have a complete specification of those conditions. A rule is not so determinate that you can specify once and for all the sorts of conditions under which to apply it. Even if you could state those conditions you still could not specify all the instances you would identify as falling within those conditions were you to confront them. For identification powers are not static. They grow and accommodate family resemblances in an unpredictable way. Thus

---

29 The relevant passages in Philosophical Investigations (Oxford: Basil Blackwell, 1972) are secs.201-41.

although an agent might now know more about the rule's application conditions, he need not be said to have changed his conception of the rule. It remains the same as the one he applied in the past.

The dilemma is real. Shall we say the rule has been changed because we hold rules to be determinate? Or is it the same because we allow a measure of indeterminacy in the conditions of applications of rules? Wittgenstein's, answer, though not always clear, is that we must decide case by case. But it is essential that we recognize that we are deciding and not discovering how things are.

What this means for a formal theory of mind is not clear. It at least suggests, however, that many of our decisions about rational relations, whether deductive, inductive or explanatory, are not decisions that are predetermined or fixed by a definite class of rules. If we feel the need to use rules to classify or systematize the class of sentences we judge as rationally related, then we must at least recognize the openness of that class. We must allow that new rules may always be introduced, or recognize that our old rules are being extended. Given this feature of 'openness', it may be unhelpful to think of our judgements of rationality to be rule-driven. Sometimes we just do recognize an inference or judgement to be rational. There is no principle we appeal to, or need to appeal to, in order to explain or justify why it is rational. We just know it is. We recognize it as rational. If someone disagrees with us, the best we can do is to show them the conditions under which the inference or judgement was made and trust that they are sufficiently like us to see its correctness. We cannot prove it to them if they disagree, however, for there is no pre-existing proof procedure to follow. Perhaps we can create one

specially for each case. But in general, these are merely teaching aids to assist us in the task of getting others to learn to recognize inferences and make judgements as we do.[30]

The upshot of Wittgenstein's comments on this point is that the ultimate grounds for human agreement in matters of rational judgement are to be found in our distinctive learning predispositions. What makes two people come to agree over the identification and classification of objects is not some reified rule which both have grasped. It is their inclination to recognize resemblance in the same sort of way.

There is an echo of this view in Nelson Goodman's discussion of induction.[31] Induction is the expectation that future cases will work out like past ones; that on the basis of observing a trait present in a subsection of a class one can generalize to the traits present in all members of the class. The problem Goodman puts forward is that for any class there are an indefinite number of predicates that we might choose to generalize, and yet which we never consider. Everything is similar to everything in some respect. Any two things share as many traits as any other two, if we are undiscriminating about what to call a trait. Yet in most cases we never even consider certain predicates for projection. Suppose that many emeralds have been examined for colour and all have been found green. Since all up to now have proved green, we expect the next emerald examined to be green. However, imagine we came across a tribe of aliens who use a different adjective for the

---

30 See, for instance, Wittgenstein, ibid. p.227.

31 Particularly, Fact, Fiction, and Forecast, (New York: Bobbs-Merill, 1965).

colour of emeralds, "grue", explained as follows: an emerald is grue if it is examined before New Year 1984, and is green, or else it is not examined by then, and is blue. Thus the grue emeralds comprise all those green ones that will have been examined by New Year 1984 plus all blue ones, if any, not examined by then. Since emeralds really are green and not blue, all emeralds examined before New Year will be green. But it is also true to say they are grue. Solely on inductive grounds both are equally likely to be true of emeralds as a class. And yet no humans actually believe that all emeralds examined after The New Year will be blue. No one believes it, but it is unnervingly difficult to say why that inference is not legitimate while the inference to greenness is. Goodman tells us certain predicates are projectable, natural. The mind finds one inference, and one type of predicate better than the other, even though in some abstract sense there is equal evidence for each. It is a fact about our minds.

But if it is a fact about our minds, is it a computational fact? Why do we share intuitions of projectability, rational inference and similarity? Both Wittgenstein and Goodman treat our dispositions in these matters as basic, pre-rational. There is no computational reason to bias one conception of similarity over another.

Might the same not be true for our sense of 'proper' belief management? Might it not be that we simply do recognize one belief set as a rational successor to another? Must there be rules here? Recognition of family resemblance is one capacity which computers have so far been lacking. The real explanation may be that the capacity

flows from the hardware up.[32] Nature designed human minds to project only certain predicates and to register as similar only certain resemblances. The raison d'être may be evolutionary, but the method is non-computational. The possibility that the control mechanism regulating our belief system operates without rules is certainly not well confirmed. And yet it has a prima facie plausibility when one recalls how difficult the updating problem is for systems with virtually unbounded background knowledge. The frame problem is a case in point. Notified of a single change in our environment, the modifications we are obliged to make to our belief system ramify from frame to frame.[33] Nature never changes in just one respect. A change in one object changes its relations to an indefinite number of others. As beings who share a vast storehold of common knowledge, we do not state the ramifications of each change explicitly; we communicate by identifying salient changes. Ramifications are unstated. But such communication can succeed only because we update our belief systems synchronously. It is too early to say whether this synchronous capacity reflects our common hardware or our common programming. But it is an hypothesis too serious to neglect.

In the next chapter I shall consider more deeply the notion of non-computational abilities and inquire whether our capacity to make sense of the actions of others -- that is, to empathize and put ourselves in other's shoes -- may not be non-computational too.

---

32 Scott Fahlman in "Representing Implicit Knowledge" has argued that only large parallel systems can handle recognition task in real time. The basic components of these systems are represented by "very simple hardware elements called nodes, relations among.... them are represented by additional hardware elements called links". p.152 in Geoffrey Hinton and James Anderson (eds.) Parallel Models of Associative Memory (Hillsdale NJ: Lawrence Erlbaum, 1981).

Chapter Five

EMPATHY

In this chapter I shall explore the prospects of formalizing rational choice. Rational choice lies at the heart of central cognitive science. It relates to deliberation, considered judgement, thought and the rest of the processes we believe distinguish human decision-making from that of animals. Rational choice is the ultimate target of cognitive science. It is the arena of 'real' thought: a domain where occurrent ideas interact with occurrent ideas and where values bear on choice. Human decision-making lies at the 'interface' of private and public. For it is in matters of action that we strive to bring public norms of justification into our private lives.

I have been arguing that the prospects for theories of <u>peripheral</u> cognitive faculties are good. The tasks of our peripheral faculties are so restricted or circumscribed that we have every good reason to hope to find knowledge level theories that explain their structure. On the other hand the prospects for knowledge level theories of <u>central</u> cognitive faculties are less sanguine. The tasks these faculties have been designed to tackle are open-ended. There are possibly an indefinite number of cognitive strategies for attacking them, some, no doubt, which we have yet to imagine. Moreover, it is doubtful whether these strategies have any claim to being 'natural' or 'efficient' in systems of different functional architectures. It was in the light of

this suspicion that I considered arguments, last chapter, about the possibility of ever finding a formalism in which to represent belief change as a formal inferential operation. By showing an inference to be formalizable we show that there may be natural joints in the process. Our formal theory serves us as what Marr called a computational theory: it shows us that a real solution to the information processing problem exists, and can be implemented. If there is more than one computational theory there may be grounds for choosing between them and so selecting at the knowledge level the 'natural' computation. Thus in proving a task to be formalizable, we show it to be open to analysis at the knowledge level and hence to be non ad hoc.

It is important to emphasize, however, that failure to find a formalism in which to represent a process, such as belief change, as formal, does not mean that we shall never have mechanical models of those processes. A.I. research in belief management is a healthy and necessary branch of cognitive science. Similarly, in cognitive psychology, research by experimentalists such as Kahneman and Tversky,[1] is uncovering widespread biases in our internal mechanisms of belief management. But we must wonder whether these models tell us much about the rational management of belief. I tried to show that we do not have a particularly clear idea of what counts as relevant knowledge in the field of rational belief management. We have no shortage of intuitions about the sort of criteria that well-managed belief systems satisfy, but these intuitions vary, and may possibly conflict. There is no unambiguous metric of rational equilibrium in beliefs: the concepts of

---

1 See especially D. Kahneman, P. Slovic and A. Tversky (eds), Judgement under Uncertainty: Heuristics and Biases. (Cambridge: Cambridge Univ. Press,1982).

coherence, consistency and explanatory plausibility are built on a set of unformalizable conventions and norms. The degree to which this is true is to be found by a detailed study of the different logics we acknowledge as having foundations in our intuitions of rational inference. Surely it is not an accident that there are so many modal logics, epistemic logics, inductive logics, deductive logics and theories of explanation.

I concluded last chapter with the hypothesis that the reason we achieve agreement in our judgements of the rationality of transitions from one belief state to another may be due to our conditioning and our biological hardware. That we march in step in innovation, extension of rules, choice of projectible predicates and so on, may be more a fact about our neurological predispositions than a fact about our computational nature. Belief modifications might be partially the product of non-computational processes.

In this chapter I shall extend these arguments to include our judgements of rationaliy in action. The practical syllogism is to practical reasoning what the deductive, inductive, and abductive syllogisms are to theoretical reasoning. We rationalize action by showing it to be the outcome of a valid practical syllogism. If all is not well in the world of theoretical reasoning, how much worse are things in the world of practical reasoning? This I intend to show this chapter. After considering how reasoning might be non-computational, I examine the form and content of reasoning directed to practical ends. Most humans have the capacity to recognize when instances of this reasoning are valid. How is this achieved? Can the ability to recognize validity in practical reasoning be formalized? Can it be

recast as a process similar to mathematical thought?

How Might Reasoning be Non-computational?

Before I begin an examination of practical inference I would like to consider how reasoning could be non-computational. Since it was the evident rule-governed nature of reasoning that led Aristotle and eventually Turing to propose formal, computational models of thought and proof, it might seem that the closer we get to matters of pure reasoning -- as in deliberation, planning, evaluation and so forth -- the greater the chance of finding formal models. No doubt, for many of the cognitive tasks we perform this is true. Much of our agreement about validity in deduction (though clearly not all), about the proper method of calculating, and so on, can be explained by pointing to the formal basis of deductive proof, calculation, etc. But there is no way of showing that agreement in thought and action, requires a formal basis. The primary reason for suspecting that some of our rational faculties might be non-computational is precisely that we can recognize rational relations without being able to formalise the basis of our recognitional capacity. Just as we can recognize family resemblance without being able to state the set of attributes or relations in virtue of which we can recognize resemblance, so it may be that we can recognize that a certain proposition set 'is made reasonable' by an antecedent set, or that some action 'makes sense' in light of an agent's particular beliefs and desires without there being a set of formal rules which justifies or proves the validity of our intuitive inference. The reason we all agree in these matters, may be found in a study of some of the lower level features of our biology. Recognition

may be too ill-structured to be a computational process.

As far as I know there is no hard evidence that recognition is a non-computational process. There is even less evidence to show that the way we do in fact understand belief change, desire change and intentional choice is in principle non-computational. But we can readily imagine why it might be so. Suppose our faculty to recognize the rationality of shifts of belief is due to some simulation faculty we have. To test the rationality of a belief change we run it through our simulator: Does it feel natural? Would we be inclined to accept it? If we test the plausibility of a rational inference by simulating the inference internally -- accepting new beliefs as rational if we think we would accept them were we to believe certain other propositions, and rejecting others as irrational if we think we could not -- then it may well be that there are no computational processes involved.

If we grant that understanding rational inference is a simulation process there are good reasons for assuming it to be non-computational. Simulation is like growth. We set the system at an initial state and let its own mechanisms of change play out. No constraint is placed on the order or quantity of causal interaction, and the whole thing unfolds as the result of local micro-interactions accumulating into macrochanges. The outcome of such local interaction may in fact satisfy certain global constraints. But there is no requirement in simulation that it must satisfy any constraint other than the one condition that the trajectory of e.g. belief system changes is one the simulator itself can produce.

The reason this is not an utterly implausible model of change in attitude management is that so many auxiliary or background beliefs must be present -- or if they are beliefs which might interfere with belief transition, then not present -- that the operation mapping the state of a belief system at t_1 into a state at t_2 may be impossibly complex. In fact it may be one-many, that is, it may not be a function at all, for there may be several equally rational transitions that are possible at any moment.

Alert to this problem, cognitive scientists have two replies. First, they remind their critics that theories of belief management which cite inference rules are not process theories, for inference rules are permissive not determinative[2] and so only state the inferences that are possible. Hence the logic of belief management defines only a vocabulary of inferential steps available to the reasoner[3]; it does not itself generate a connected chain reasoning, a trajectory of actual belief changes. Second, cognitive scientists argue that it is possible to put boundaries on the scope of potentially relevant beliefs. Most often this involves showing how reasoners might compartmentalize information into frames. Both face formidable challenges. If inference rules are permissive it must be possible to show that they are consistent and complete -- an issue I suggested was extremely difficult to resolve in the case of belief management. If

---

2 Such a view, for instance, has been advanced by Martin and Braine, in "On the Relation Between the Natural Logic of Reasoning and Standard Logic." Psychological Review Vol 85 #1, p. 4.

3 The distinction has been discussed by Henry Kyburg, "Functional Architecture and Free Will." in Behavioural and Brain Sciences, (1980), 3, p. 144.

information is bounded, then there must be principles for deciding what to include in frames and what to ignore. This problem, the problem of circumscription, remains unsolved. Moreover, in cognitive science, it is customary to demand more than just principles: one wants process models: the actual control mechanisms regulating the order of rule use. But is it likely that there is a logic to inference rule selection? Or a logic to framing, to compartmentalizing potentially relevant information? If we doubt that these mechanisms are principled then they reveal more about the mechanics of our brains than the mechanisms of our minds. This, at any rate, is a possibility we must not dismiss a priori.

David Marr in a paper entitled 'Artifical Intelligence -- A Personal View'[4] distinguished between processes which involve the interaction of a great many distinct causal ingredients "whose interactions are their simplest description", and processes that involve a small number of causal ingredients which can be described in fairly simple terms -- for instance, in terms of automata which operate according to a small number of machine rules.

As an example of the first sort of process, which Marr called Type 2 processes, he cited the folding of a protein. He pointed out that research has shown that when a protein folds

> A large number of influences act on a large polypeptide chain as it flaps and flails in a medium. At each moment only a few of the possible interactions will be important, but the importance of those few is decisive. Attempts to construct a simplified theory must ignore some interactions; but if most interactions are crucial at some stage during the folding, a simplified theory will prove inadequate. Interestingly, the most promising studies of protein folding

---

4 Artificial Intelligence 9; pp. 37 - 48.

> are currently those that take a brute force approach, setting up a rather detailed model of the amino acids, the geometry associated with their sequence, hydrophobic interactions with the circumambient fluid, random thermal perturbations etc., and letting the whole set of processes run until a stable configuration is achieved.[5]

Apparently to simulate protein folding one does not look for a level of analysis where the process can be represented as a simple automaton. The complexity is ineliminable.

As an example of a non-type 2 process, Marr cited the process occurring when a system is described as performing a Fourier transformation. Mathematicians have shown that there are several algorithms for performing such transformations. Systems which follow one algorithm obviously realize very different processes than ones following other algorithms. Yet in virtue of the general theory of Fourier transformation all such systems must do certain things. We can specify at an abstract level what is going on in such systems. They must perform in certain ways. In specifying these performance conditions, we show the process to be what is called, Type 1. For there is an economical characterization of the process which proves that there are non ad hoc rules for performing the transformation.

Now the idea I am considering is that there really is no Type 1 theory of what the brain or mind does when we change belief, etc.. Rational processes are in fact Type 2 processes. We talk as if there are neat laws and decomposable processes underlying belief change, etc., encouraging the idea that there is a level of analysis at which the neural processes underpinning belief change can be viewed as Type 1

---

5 ibid. p.39.

processes. But this may just be talk. Our capacity to recognize rational change need not presuppose it. Some cognitive processes are irreducibly complex.

Marr was not the first to distinguish irreducibly complex processes from neatly decomposable ones. The suggestion that there are processes in nature that are so complex that there is no way to economically describe them in some formalism, comes originally from von Neumann.[6] Von Neumann had been considering the import of McCulloch and Pitts'[7] discovery that any perceptual or cognitive process carried out by a living nervous system which could be precisely defined could be logically simulated by a network of abstract neural modules of considerably simpler structure than living neurons. He questioned whether their proof showed that there is in principle a computational theory for every cognitive or perceptual process.

In one sense, it clearly did. For if one could specify exactly what the brain was doing in 'seeing' a square, or in 'calculating' the product of 7^4, (or 'judging' a belief to be rational), then a neural model could be designed to simulate that process. The trouble was, however, that although

> the insight that a formal neuron network can do anything which you can describe in words is a very important insight and simplifies matters enormously at low complication levels, it is by no means

---

6 The Computer and the Brain, New Haven, 1958, and Theory of Self-Reproducing Automata, edited and completed by A.W. Burks, Urbana, 1966.

7 "A logical calculus of the ideas immanent in nervous activity." Bulletin of Mathematical Biophysics, 1943, 5, pp. 115-133.

> certain that it is a simplification on high complication levels.[8]

If a large proportion of the brain were involved in a process, as is the case when we recognize two reasonably complex shapes as similar, then it is not clear a priori that there is a simpler description of what the brain is doing than that given by describing the brain processes themselves. As he put it

> Normally a literary description of what an automation is supposed to do is simpler than the complete diagram of the automaton. It is not true a priori that this will always be so. There is a good deal in formal logics to indicate that the description of the function of an automaton is simpler than the automaton itself, as long as the automaton is not very complicated, but when you get to high complications, the actual object is simpler than the literary description.[9]

Von Neumann was skeptical of finding Type 1 descriptions of many cognitive processes; he seriously entertained the idea that much of a our cognitive life is the outcome of irreducibly complex neural processes.

Such a view is tantamount to denying that there is a 'real' level of analysis above the neurophysiological (or possibly low level computational, where the interaction of the neurons themselves is explained in automata terms) which effectively describes the processes going on in the brain. The belief of the cognitive scientist is that there is a robust level, the knowledge level, or failing that, the

---

8 Von Neumann, 1966, op. cit. as quoted in Robert E. Shaw, "Cognition Simulation and the Problem of Complexity", Journal of Structural Learning, 1971, 2(4), 31-44, p. 153.

9 ibid. p. 154.

representational level, which is related to the neurosciences, roughly as biology is to biochemistry, or software to hardware, or to use a metaphor I used earlier, as catastrophe theory, as exploited in causal embryology, is to biochemical growth, and which can be characterized in the language of automata theory. If certain cognitive processes are Type 2 processes, this belief is false. There is no simple characterization of what is going on in the brain. Consequently, the apparent simplicity of our talk about our beliefs and desires, and the ease with which we grasp rational relations between them, would not be due to computational processes underpinning belief and desire change, and rational choice. Rather, talk of belief and desire would be a means of triggering in us Type 2 processes that effectively simulate the Type 2 processes in other agents. They would be aids that allow us to recognize patterns of 'rational' order in processes; but this order would not, in any sense, be mathematically simple or well-defined. We would understand others through empathy. But empathy would be basically a non-computational process.

I shall now try to show that the cognitive processes underpinning our judgement of the rationality of actions, as displayed in our explanations of actions,may be Type 2. Our belief and desire language may be a deceptively simple way to refer to immensely complex Type 2 processes. If we can show that rational explanations of action cannot be translatable into mechanistic explanations, then the mere presence of an apparently simple explanatory schema, the practical syllogism, is no evidence that the actual processes causally responsible for rational choice and for understanding rational choice, i.e. empathy, decomposes neatly. Philosophers have suspected that because we can represent

rational choice in the practical syllogism that therefore there must be an effective procedure for deciding whether the conclusion of the syllogism follows from the premises. They have assumed that there are a set of formal rules that establish the validity of the syllogism. These form the core of the process of rational judgement and, therefore, of rational explanation. Hence people agree in their judgements about which actions 'make sense' because they share a rule governed competence to judge the rationality of actions. Just as in linguistics, our judgements of well-formedness are supposed to derive from our knowledge of a set of grammatical rules, which determine all and only syntactic structures of a language, so in practical inference, our judgements are supposed to derive from our knowledge of the rules of rationality, which determine all and only actions that are rational given certain antecedent conditions.

In what follows I shall challenge this view. I shall argue that the capacity to see A as a rational thing to do in circumstances C is non-formalisable. I begin with a general discussion of the ingredients in rational explanation, noting that, as it stands, rational explanations are enthymematic: they are missing premises which they suppose the reasoner can fill in. When it comes to stating just what these missing premises are, however, we discover a tangle of difficulties.

Explaining Action

An intentional action is one which is undertaken, consciously or unconsciously, for a reason which one believes applies in one's current situation. Philosophers typically distinguish reasons themselves from

beliefs about reasons: the distinction being that reasons can exist without an agent knowing or believing them. When it comes to explanation, though, it is belief about reasons that counts. An agent who does not recognize connections between certain facts and a state he desires will remain unmoved by those facts. He may have the means at hand to achieve an end which, all things considered, would be one he would be advised to pursue; yet his failure to appreciate the readiness of those means, and his failure to recognize them as means to his ends, implies that he cannot act for that reason. Whatever his action, it is not to be explained by citing that reason. A reason turns no gears in a rational machine, unless it is grasped. And for that, the agent needs the capacity to recognize what it is for something 'to be a reason'. Evidently rational explanation presupposes that agents have the means of recognizing reasons. The only issue in doubt is whether this means is formalizable; is it the result of tacitly following an effective procedure?

There are any number of normative formalisms, some complete, most partial, which we appeal to time and again to make sense of thought and action and which might therefore prove the formal basis of practical inference. But, in fact, they all presuppose the ability to recognize reasons. For instance, when we see an action, whether one performed by ourselves or by someone else, as rational, we see it as 'appropriate' to the situation. Given the context of action we recognize that 'it makes sense', it is a 'reasonable' thing to do. We are not told how to see the rationale of an act, to grasp its point. It is assumed that we have the ability to understand actions as 'making sense'. Sometimes the explanation proceeds by idealizing the situation. Either it is

simply assumed that we have an idea of how an ideal agent would perform, under certain idealized and simplified circumstances, or we are given an explicit account. But there are no specific rules telling us how to understand the ways ordinary people fall short of the ideal.

Another method of explaining action appeals to an even less explicit method of assessing rationality. In Verstehen explanations, we describe the salient features of a situation as completely as possible and rely on our 'intuitive' capacity to grasp what counts as reasonable in order to see an action as plausible and hence as having been explained. As long as we all operate with similar capacities, similar notions of 'reasonableness', we agree in our judgements of what actions are rationally explicable. But no attempt is made to explain the basis of these capacities.

Yet a third method of explaining action is to describe first, the image an agent has of himself, and then the situation he takes himself to be in. As we increase our grasp of 'scripts' which people live by, we are able to recognize a wider and wider circle of actions as 'consonant' with roles people think they are playing; we can see more actions as 'cohering' with certain characters; as 'conforming' to certain aesthetic standards characteristic of certain roles. Once again, though, the explanation provides no explicit account of the rules for interpreting coherence, conformity or consonance. The ability to recognize particular actions as coherent, etc., in the light of given scripts, is assumed basic to all scientists using this explanatory schema.

All such explanations appear to have a common structure. They begin with a description of a situation S, which an agent A, is

supposed to believe. They cite an end, goal or set of desiderata D, the agent wishes to achieve. And they conclude with an action, a_i, that seems reasonable to the agent, given his grasp of the situation and his desiderata. Schematically we may represent this as in PR below.

PR

P1: A believes he is in situation $S(s_1, \ldots. s_n)$

P2: A operates with desiderata D

C: therefore A believes a_i is a reasonable act to perform

A's actual performance of a_i is then explained by attributing to him a <u>disposition</u> to do what he believes is reasonable, other things being equal.

Now the inference from P1 and P2 to C, and which lies at the heart of the explanation, is sanctioned by the reader or agent if it seems to him that D and S do, in fact, render action a_i reasonable. There is no escaping the psychological character of this process of sanctioning, this capacity to recognize reasonableness. Just as the notion of a proof, until the last years of the 19th century, was primarily psychological: an intellectual activity that aimed at <u>convincing</u> oneself and others of the truth of the sentence under discussion by means of intuitively convincing arguments; so, the practical syllogism, until recently, has been a means of intuitively convincing oneself and others of the reasonableness of a given action. The question is: can this capacity to recognize arguments as intuitively convincing be explained in terms of a few formal rules operating on descriptions of D and S? Can we provide a formal mechanism by means of which we can validly infer from D and S just those descriptions of acts that we do in fact recognize as reasonable?

Attempts to formalize PR begin by inserting intermediate steps between P1, P2 and C. It is obvious, as it stands, PR is enthymematic. Thus between P1 and P2 it is customary to insert a clause about the action the agent sees open to him in his current situation. Because P1 merely states that A has beliefs about the current state of his world, it is quite possible that he has no further beliefs about the 'affordances' of the situation -- the obvious availabilities it offers him in terms of action. Since choice is defined over actions we require that the agent have beliefs about his action. These must be made explicit. Hence after P1 we have

P2.1: A believes in S he can do $(a_1 \ldots a_n)$

Next, to guarantee that A brings his desiderata to bear on his action set, we must state the consequence of each action and the value of each consequence. Most formalisms of practical reasoning depend rather heavily on some form of tacit means-end reasoning. It is because we believe that a certain action has effects on the world which bring us closer to our ends, or to states which we value, that we undertake actions. It may be objected that this consequentialist approach is overly restrictive. Not all practical inferences rely on means-end reasoning. For instance it does not seem to accommodate the Kantian who thinks that actions should be undertaken irrespective of their consequences. A lie, Kant said, should never be told, whatever its consequences. Its identity suffices to settle questions of choice.

Yet even Kantians admit that the rationale for living in accordance with formal rules of morality is to be found in the logical

consequences of violating them. One binds oneself to the institution of truth-telling because of the undesirability of a world lacking the institution. In one sense, then, even Kant accepted that one chooses to speak truly to insure some end, it is just that the end he selected was impersonal: the maintenance of an institution. Accordingly it may not be unreasonable if we assume means-end reasoning to dominate practical reasoning. To make the means-end elements of our practical reasoning explicit, it is necessary to add two further premises:

P3: A believes of any action a_i which he considers performing that it has end-relevant consequences $(s_i,...s_j)$;

P4: A assigns a value to every $(s_i,...s_j)$ according to his desiderata D.

Only one further premise remains to bring the process close to mechanization. Since P1 to P4 entail a conclusion about choice only if there is some rule for deciding between $a_1,...a_n$, a final premise concerning decision rules must be added. An obvious candidate is:

P6: A prefers whichever action a_i has the most valued consequences.

Thus we have:

PR^*

P1: A believes he is in situation $S(s_1,...s_n)$

P2: A believes in S he can do $(a_1,...a_n)$

P3: A believes of each a_i that it has ends-relevant consequences$(s_1,...s_n)$ or probable consequences $(ps_1,...ps_j)$

P4: A operates with desiderata D

P5: A assigns a value to each $(s_1,...s_j)$

P6: A prefers whichever action a_i has the most valued consequences.

C : Therefore A believes a_i is a reasonable act to perform.

It must be apparent that by a simple expansion of common-sense reasoning we have come close to a reconstruction of reasoning as laid out in the modern theory of rational choice: a theory that was formalized as part of the foundations of economic theory. But the two are different in interesting ways. It may be worth comparing them.

According to the formal theory of rational choice, an action is provably rational, only if we can identify:

1. A set $A(a_1,...a_k,...a_o)$ of behavioural alternatives 'objectively' open to the agent.

2. A subset $A^*(a_1,...a_n)$ of behavioural alternatives which the organism 'considers', or otherwise regards, as feasible,

3. A set $S(s_1,...s_n)$ of possible consequences, or states of affairs, that represent the possible outcomes of choice.

4. A 'payoff' or 'utility' function, v(S), representing the value placed by the organism upon <u>each</u> of the possible oucomes of choice.[10]

5. Information concerning the probability that a particular consequence will arise if a particular alternative a_i in A or A^* is chosen. The information must be complete enough to associate with each element s_i in the set S a probability $P(si)$ which is the probability that s_i will occur if a_i is chosen. This information constitutes a theory of the environment.

6. A decision rule, R, which ensures that the agent selects a single action from the set A^* which he considers. For example: a function R taking the expected value of each a_i in A^* and selecting one as the most rational on the basis of some property it has, such as the highest expected utility.

Rational choice is then seen as the end product of two successive filtering devices. The first is defined by a set of constraints which cuts down the possible set of courses of action and reduces it to a vastly smaller subset of 'feasible' actions. The constraints are assumed to be given and not within the control of the agent. For

---

10 In some models of natural choice U(s) orders pairs of elements of s without assigning a cardinal measure to their value.

instance, the agent considers only those actions it has the power to physically, technologically or economically undertake. The second filtering process is the mechanism that singles out which members of the feasible set shall be chosen. It relies on information about the probability and value of consequences in order to isolate one action as the best. Theories of rational choice assert that this mechanism is designed to maximize some objective function, be it a real one like profit, which is objectively measurable in the external environment, or a purely notional one like expected utility which is a subjective property of actions as conceived.

We may represent the decision making process as given by the formal theory of rational choice as in Fig 5.1.

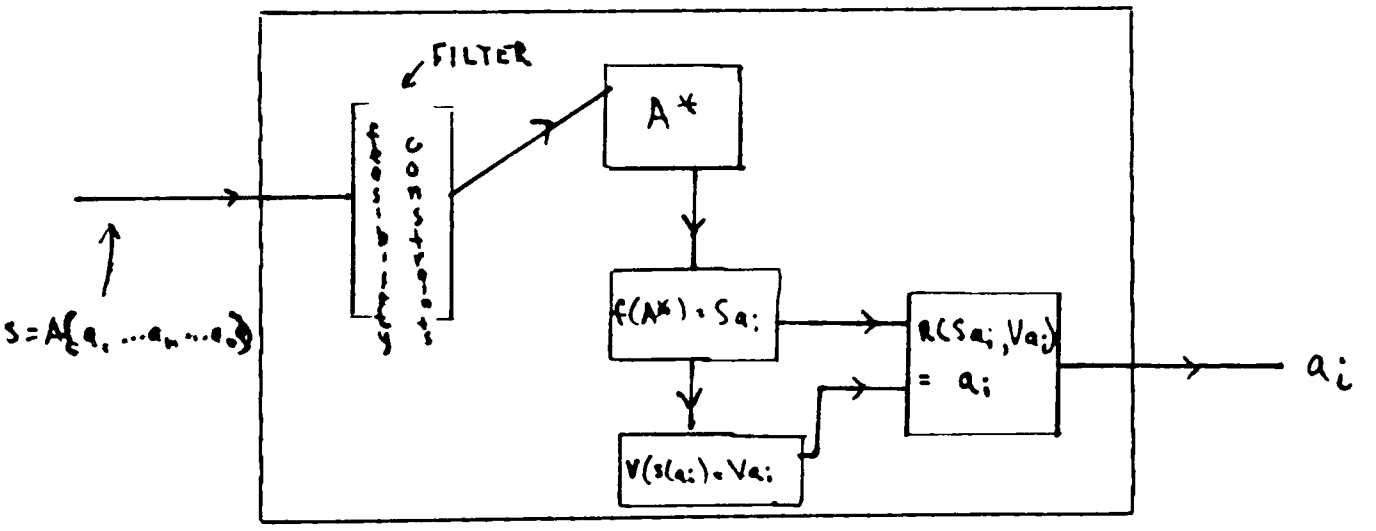


Figure 5.1

Have we filled in the missing steps in the practical syllogism in sufficient detail to warrant seeing it as a formal reconstruction of the process of practical reasoning? To be sure, the theory of rational choice does not pretend to be an account of the actual processes occurring during practical inference. It is most unlikely that to decide how to act, we actually form beliefs about the

consequences of each of the actions we implicitly believe we could perform. Our thinking abilities are modest compared to the complexities of our environment. Being endowed with limited cognitive resources we haven't the computational power, memory capacity or speed of thought to generate and test each possible action. It is more efficient for us to work with an impoverished conception of what is possible, often ignoring consequences of feasible actions that are far down the chain of cause and effect and which are conditional on hundreds of other events occurring. Hence the formalism we are comparing should be thought of as a rational reconstruction of the process of practical reasoning. Better yet, it should be thought of as a 'proof' that the process of rational choice could be mechanized in a principled and non ad hoc manner. Nonetheless we do require that the reconstruction capture the critical moves in rational choice. Thus, just as in syntactic theory, we assume there is a core syntactic competence which incorporates a certain body of knowledge that regulates syntactic processing despite the competence being embedded in a larger system of processes, so we are supposed to be able to see the rational decision theory model of practical reasoning as identifying a core competence to think practically, despite its also being embedded in a larger system of processes.

Now, on analogy with linguistic processing, this core competence to reason practically ought to depend on knowledge that is specific to practical inference. This cannot be a necessary condition, for the body of knowledge regulating each of our cognitive faculties can to some degree be affected by changes in the beliefs and decisions of the system as a whole. But each body of knowledge should be relatively

complete with regard to the task environment to which it is related.

To the extent, therefore, that rational decision theory offers an effective account of the knowledge that is relevant to practical inference, it provides a knowledge level theory of the sorts of knowledge based processes involved in practical reasoning. To the extent that this account is defective, incomplete, or unrealistic with regard to the task of reasoning practically, it falls short of an adequate knowledge level theory of practical reasoning. I want to show now the ways in which it falls short.

Problems with Rational Decision Theory

To appreciate the full complexity of the formal model of rational choice we must look at the sorts of knowledge it invokes in the course of computing a choice. These are of three basic types. First, there is knowledge of the constraints which serve to filter the class of 'objectively' or 'environmentally' possible actions to ones that are feasible. These may be as simple as budgetary constraints or as complex as the constraints which flow from a moral theory, outlawing some actions as beyond the moral pale. There is second, knowledge of the causal consequences which performing particular actions may be expected to produce. Relative to a particular world, where certain causal laws operate, actions have predictable consequences, unless there are exogenous interventions. To calculate the consequence functions an agent must have a theory of his world and a theory of the background and boundary conditions of that world. Third, the agent must know what he values and how to apply those values to particular events so as to order the desirability of all possible events in his

world. To be able to apply a utility function to all events, an agent must know not only how to evaluate states, but how to evaluate the importance of his normative criteria. Where conflict arises due to the use of more than one norm, the agent must know how to resolve it. He must know rules for conflict resolution.

Philosophers and social scientists have often charged the formal model of rational choice with being an extravagant act of idealization. Agents have neither the computational power nor the knowledge of causal law to derive the possible consequence set of each of their possible actions. Nor are they able to assign a utility measure to each potential consequence, and then arrange the set in a strict preference ordering. Such extravagances may be tolerated, we are told, because the formal theory is describing an idealized competence. People fall short of these ideals in ways that are meaningful, hence to investigate practical reasoning we should import the rational choice schema and study deviations. Patterns will emerge that are instructive about the actual representations and control mechanisms the agents are using. All these patterns would have been invisible had we not disciplined our thinking in accordance with the model provided by rational decision theory.

Here we have the standard justification for approaching practical reasoning at the knowledge level. It assumes that there is a single faculty of practical thought; that there is a univocal notion of rationality underpinning our capacity to recognize practical inference; and that the formalism of rational decision theory is an effective way of representing this faculty. Moreover, it assumes that each of the knowledge bases invoked in the formalism can be articulated clearly and

evaluated for adequacy. Unless the consequence function, the utility function, the decision rule function and filtering function are well-defined, the formalism will be empty. For a formalism that depends on the application of several functions which are ill-defined is itself ill-defined. Adherents of the formal theory assume such functions are indeed well-defined. But when we look closely at what is required to compute these functions we shall see that there is less precision than their simple mathematical formulation reveals. This is what I propose to show.

The Consequence Function

In the formalism of rational decision theory the consequence function assigns to each possible world state an estimation of its probability given a particular action. It is a mapping from an ordered pair of action and world state (a_i, s_j), onto a probability measure. In our reconstruction of common sense inference the job performed by the consequence function is done by means-end reasoning. What we must decide now, is whether this reasoning computes a function.

In means-end reasoning we are, in effect, engaging in causal analysis. We know the situation S we are in, we infer the possibilities of action open to us, and we spin out, on the basis of knowledge of laws, of background conditions and of boundary conditions the probable effects of undertaking arbitrary actions. If the world (environment of action) is more or less closed to outside influences, and if the interactions which occur within it are functions of a finitely enumerable set of interacting properties, then as long as we can be confident that there are no counteracting forces to disturb or

interfere with the normal causal dependencies, we can state with some assurance how probable it is that particular actions will bring about certain consequences. It is a condition on such explanations, though, that they are reliable only to the extent that we are justified in making our assumptions about system closure, boundedness of interactions and normality of systemic operations. If these assumptions are justified, our causal analysis will be justified. If they are sound and complete, the process of causal reasoning may be regarded as computing a function.

Now, in rational decision theory, because the environment of action is construed as a system of affordances, each situation is defined by the class of actions it permits. We do not define a situation as a world state, as we do in common sense reasoning. Rather situations are defined by their potentials for action. Consequences, therefore, are not determined by causal reasoning as normally practiced. We do not take a world state and apply a set of laws to it in order to determine the world state at a future time, other things being equal. For since situations are defined by the potential actions they afford the agent, we would require laws that express nomological connections between sets of affordances. It is true, such laws might be found for closed worlds, where the only transformations that can occur are the result of performing one or another of the actions afforded the agent. Thus, in a game of chess, since the only relevant state changes are those defined by legal moves, we can define each situation as a class of potential moves, and consider laws or strategic rules defined over those situations. The effects of a given move, therefore, might be described by the sorts of strategic situations it could give rise to.

But this approach has unreasonably constrictive limitations. The world of action is not as narrow as the world of chess. We live in a spatio-temporal world where the agency of change is frequently natural law. Thus we cannot predict all the future consequences of an action unless we also know the world state associated with a situation. Natural laws are defined over states or properties, not over sets of action affordances. Moreover, in the real world, the potential for action which a given situation offers us, is not as clearly delineated as we find in chess. Games have rules determining the class of allowable moves, but in life, no such rules exist. Thus, although we live under constraints of norms, social conventions, budgets and the like, few of the environments we operate in are so rigid that we cannot <u>improvise</u> new modes of conduct. All things considered, a decision making system must have access to knowledge about the manifest properties of the situation it is in. It must know more than just what actions are open to it at any point in its environment. Otherwise it will be unable to create a consequence function of any use in non-artificial situations.

Yet as soon as we try to state what extra knowledge a system must have to be able to determine its consequence function in real world contexts we stumble across a familiar problem. If the consequence function is to apply to real world situations it must have access to information about causal laws, boundary conditions, normal conditions and so on. Only if these can be shown to be sound and complete, can we be confident that a system may operate with a consequence function that generates unique consequence sets. If we believe -- as I shall argue we should -- that there is an essential openness or indeterminacy to

the boundaries of the environment of action (that is, the world of relevant cause and effect), or, if we believe that there is an openness or indeterminacy to the background conditions of the environment, making it nearly impossible to state exactly what constitutes the normal conditions of that environment, then we will have good grounds for suspecting that the consequence function is ill-defined. Accordingly, unless we can state with some confidence that the states of the world $S(s_1,...s_n)$ over which the consequence function is officially defined are <u>all</u> the end-relevant states of the world, we will have little reason to think that the formal theory of rational choice canvasses enough of the relevant consequences of choice to identify a <u>rational</u> choice.

For example, imagine that we are standing in front of a large buffet dinner, about to choose our first course. The situation may be described in terms of the choices open to us: we can take prawns or jellied eels, smoked salmon or salad... . Or, it may be described in terms of the occurrent world states: there are prawns of limitless number, one whole smoked salmon... three plates, two knives and forks, plus endless descriptions of the decor, other guests, atmosphere, and so on, and so forth. Because real world situations contain countless properties that are irrelevant to our ends -- properties that seem to us to lack <u>normative dimensions</u> -- the art of rational decision making consists in bounding the system of properties that can, in the normal course of affairs, be affected by our actions so as to contribute, in one way or another, to the desirability of those actions. To continue with our example, our natural inclination is to assume that the normative dimensions of choice at a buffet dinner, will be confined to

the properties of the food which will have consequences for our health, our gustatory pleasure, our sense of culinary style and the like.

Yet suppose we are told that the plates on which we are to collect our food are porous and allow liquids to seep through. We suddenly have new consequences to ponder: which food is least likely to liquefy? If we are told further, that we are to hold the plates in our laps, new desiderata enter the picture. In determining the consequences of our actions we <u>assume</u> countless features about the situation. To keep the decision problems tractable we lump such features into catch-all phrases such as "providing conditions are normal", or "assuming other things are equal". What we mean is that if conditions are 'normal' then the consequences <u>relevant to our ends</u> will be... Yet it is notoriously difficult to specify what conditions count as normal since these not only vary from occasion to occasion but themselves depend on our ends. If we were indifferent to untidiness or to mess, for example, would it matter whether our plates leak, or whether we eat from our laps? Such qualities of the circumstance could be ignored as normatively irrelevant; they could be treated as part of the normal conditions alluded to in the "other things being equal" clause, features of a situation which have no role to play in cost-effective decision making. Hence, to define what falls within the class of relevant conditions, that is, to itemize the properties we must be aware of in our decision making, we must refer to the ends which those properties might bear on. Only relative to ends can we set boundaries on the causal relations that must be identified by the consequence set.

It is here that the real trouble begins. Why assume that our ends form a closed, complete and consistent system of the sort required to allow us to define a determinate consequence function? Due to the logical priority of ends over means just shown, we have been forced to shift our attention from questions about the determinacy of our consequence functions to prior questions of the determinacy of our ends. Before we can define the boundaries within which the consequence function must operate, we must identify the ends that are to be acknowledged as occurrent. Are the prospects good that these ends will themselves comprise a well-defined set? If they do not, then we cannot define the universe of relevant consequences. Let us look, therefore, at the determinacy of ends.

The Openness of Ends

One thing which geometry, chess, vision and grammatical speech have in common is narrowness of ends. The ends of the geometer is geometric proof, of the chess player, checkmate, of the seer, veridical vision, and of the speaker, grammatical speech. With such circumscribed ends it is no surprise there are circumscribed consequence functions. Central tasks are characteristically not so circumscribed. That is the problem they pose: given the openness, indefiniteness and unforeseeability of practical tasks, attempts at codifying some 'logic' of central tasks are bound to distort realities by falsely representing what is essentially open as what is susceptible to closure. As Wiggins wrote in a discussion of Aristotle's theory of deliberation and practical reason:

> No theory, if it is to recapitulate or reconstruct practical reasoning even as well as mathematical logic recapitulates or reconstructs the actual experience of conducting or exploring deductive argument, can treat the concerns which an agent brings to any situation as a closed, complete, consistent system. For it is the essence of these concerns to make competing and inconsistent claims. This is a mark not of human irrationality, but of human rationality in the face of the plurality of human goods.
>
> The weight of the claims represented by these concerns is not necessarily fixed in advance. Nor need the concerns be hierarchically ordered. A man's reflection on a new situation which confronts him may disrupt such ordering and fixity as had previously existed, and bring a change in his evolving conception of the point, or the several or many points, of living or acting. Any revealed preference theorist or other psychophysicist who seeks by extrapolation to 'take in the slack' which has been left by such indeterminacy or indecisiveness as this, or to tidy up as an 'inconsistency' the plurality of the concerns mentioned [above], is preparing to deprive his subject of autonomy, and to give him what he most likely wants not to have.[11]

Wiggin's target is the sanity of the ideal, held by some, of a system of desiderata for choice that is axiomatizable. In deliberation, we have the power to choose our ends, within restrictions. We can, with effort, perceive the normative dimensions of a situation in a fresh light. Were our value structure and ideals closed to change once and for all, there would be no scope for deliberation about ends. What point would there be in questioning accepted values or in developing standards of rational inquiry into norms, if each of us already possessed a closed, internally consistent value system? Our desiderata must be open-ended. But if that is so the consequent function must be open-ended too.

---

11 "Deliberation and Practical Reason", reprinted in J.Raz (ed), Practical Reasoning, Oxford, 1978, p. 145.

It was Aristotle's view that the paramount problem of practical choice lies in choosing the salient characteristics of one's situation. If one could identify the situation, reason might indeed determine the rational act. But how is one to identify what is salient? Dinner concludes. My neighbour lights his cigarette. I see the situation as an opportunity to smoke. Yet I might also have seen the situation as an opportunity to break an undesirable habit. In the first interpretation, my focus was on an specific act, its pleasures, its consequences. In the second, my concern would have been with my habits, their worthiness, their consequences. Can we choose which interpretation we ought to hold?

Or again, suppose we accept the dramaturgical analogy and we conceive of our actions as ingredients in the life of a dramatis persona. Is our script determinate? Are we obliged to act in routine predictable ways? Or may we improvise/interpret our part in a novel, creative fashion? What makes life interesting is variety of interpretation. There is no unique concept of a 'good' move in the game of life, nor a universal idea of a part well-played. Our parts are open-textured, being developed as we go along. There is no simple way of avoiding the responsibility of personal interpretation. Yet all formalizable conceptions of practical inference presuppose that the utility function, the desiderata, of rational agents are well-ordered and complete. If one denies at the outset that we ever finish our inquiry into value, then exposure to more of life's situations can press us to formulate more norms. We may acquire preferences for certain desires, preferences for ways of seeing the world, and preferences for ways of resolving normative conflicts. These higher

levels of desire, structure our desiderata in non-simple ways. They allow us to change our norms; indeed, they encourage us to change. Yet if our values change over time, and in step with our value changes, there occur changes in the way we perceive and conceive of decision situations, then any calculation of the consequences of an action must acknowledge that our interpretation of consequences may change too. Is it rational for us to try to build, right into our consequence function, some means of accommodating interpretational shifts? Our stand on this question determines our attitude to the formalization of rational choice. If we deny, as I think we should, that the consequences of an act are laid down, once and for all time, and so are not susceptible, in principle, of being represented by a determinate consequence function, we reject the very framework in which formal choice unfolds. We may still accept that deliberation involves the weighing of alternatives. But we should give up the naive hope of formalizing this process of weighing. The considerations that bear on understanding, or on appreciating, the normative dimensions of situations are more polymorphic than can be captured in a simple model. There is no simple consequence function to be had: no simple utility function.

Our Biology Shows Through

I have been at pains to show that underpinning our practical reasoning there is a myriad of complex processes invoking knowledge of different types, that resists, and resists in principle, being captured in a single organizational schema. The essence of rational decision theory is that if the world is not well-defined, then we must structure each

of our environments so that it seems so. The hallmark of goodness of structure, in this case, is that each environment can be represented as a complete and consistent tree of possible paths and consequences. Our choice of a particular path, or course of action, then would give rise to a determinate packet of consequences. There would never be uncertainty as to whether a particular event qualifies as a consequence, or as a probable consequence of a choice; and no pair of incompatible consequences would ever be assigned probabilities whose sum exceeds one. At the same time, our subjective evaluation of consequences would have to be complete, consistent and stable. Any state would be assessable for its desirability; assessments would betransitive, assymetric and non-reflexive; and perhaps most interestingly of all, the act of assessment would not change our evaluating function, or affect our perception of the consequences available in a situation. The resulting evaluated consequences could then be compared according to some decision rule, which would determine from a filtered set of actions -- that is, from the feasible set -- a single action or action set which would represent the rational choice.

I have suggested that the formal apparatus required to achieve this structuring does not reflect much that is true about the actual process of practical reasoning. People do, no doubt, make choices by considering options and evaluating their consequences. But much of their reasoning power is spent in setting up the framework in which to evaluate options. The interesting thing about rational choice is that people can recognize what is 'salient' about a situation. This, I suspect, reflects more about our biology than about our rationality. When a person explains an action by saying "I really had no choice.

He drew a knife. It was him or me." we know the sort of situation he is describing, the way things normally run. In short, we understand his reason. How do we manage to achieve this remarkable empathy?

The mystery, I believe, dwells in our capacity to bound or frame the situation. Somehow, by having lived in the world, we have come to an understanding of the normal course of events. We know how the human drama unfolds in worlds like ours. One day more about this process of bounding the world will be grasped. But it will be achieved only by investigating the web of human desire. What is characteristic about this web is that it seems natural to us. Desires harmonize. Yet our perceptions of these harmonies may depend on basic biological facts.

The issue here, as so often in situations which involve the management and regulation of attitudes, concerns the non-monotonic nature of our sense of rationality. In monotonic reasoning, an added premise can never invalidate conclusions that have already been reached: what follows from a set of premises still follows if the premises are added to. But judgements of rationality can become invalid from an addition to the stock of premises; for the added premise may express a condition which interferes with the inference. A judgement that some act is rational in certain circumstances, may be rescinded or annulled when we are informed of additional attributes of the circumstances. A policy that is endorsable because it secures fulfilment of a set of ends currently acknowledged, may be incompatible with a novel end expressed in a new premise. In this way, practical reasoning, unlike mathematical reasoning, is as lawyers say, defeasible.

Now, to succeed in making inferences in a world which is constantly open to new interpretation and outside interferences, the agent must have a sense of when to expect new interpretations and when to expect new interferences to arise. He must have a sense of the pattern of desire; of the sorts of ends which various situations tend to trigger in people and the sorts of ways these can be frustrated. I have suggested that this situational appreciation of ends, this ability we all have to recognize a set of ends as relevant to a situation, is not formalizable. As with other recognitional capacities, it is something we have mastered, to a more or less advanced degree, through being biological creatures of a certain sort who have been thrown in the world. How Martians might react if in the same situation may depend on Martian biology and history. There is no formal theory we can appeal to that will predict rational conduct.

At bottom, the reason lies, I believe, in the non-univocal meaning of rationality. Our capacity to see reasons as legitimate and, moreover, as compelling, may be the outcome of type 2 processes. Consequently, the understanding of reasons, would not form a single natural category of process; there need be no single faculty supporting our capacity to reason practically. Hence there may be no principled approach to practical reason.

None of the arguments presented here ought to suggest that research in A.I., or in cognitive science, will never succeed in finding mechanisms underpinning belief management, rational choice andeven the regulation of desire. To the extent to which they are sound, however, they do provide a reason for doubting that there is a single mechanism controlling our change of attitudes, and choice of action. Given the

richness of the human world, life unfolds in an open-ended space. Inquiry and human frailty force us to set boundaries on our appreciation of nature, but these boundaries may shift and change. Above all they bear the stamp of our biology and experience. In this they are ad-hoc.

CONCLUSION

I have been defending in this essay a view of cognitive science that has been current in some quarters for almost two decades, but which has recently come under increasingly heavy attack. I suggested that our primary hope for a principled approach to cognitive science lies in a two-tiered research program. On the one level, research must focus on the environment of action, searching for general characterizations of the constraints on task performance implicit in task structure, and explicating the knowledge of task structures that competent agents are obliged to have. On the other level, research should focus on the standard human mechanisms of computation and mental processing, in an effort to expose the characteristics of the human 'base' machine. Given a clear idea of the gross features of the 'base' machine, studies at the knowledge level should offer useful top-down advice about the nature of representational and control mechanisms regulating behaviour.

Moreover, it is from research at the knowledge level that we gain confidence that there could be principled mechanisms operating at lower levels. The great problem of cognitive science so far has been its ad hocness. Theories of representational form and control structures have been postulated before we know how general those theories ought to be. All too often theories are postulated to explain specific phenomena, leaving questions of generality unaddressed. The advantage of standing back and doing knowledge level research first, before struggling to

provide a fully mechanistic account with all its kluges and ad hoc elements, is that we can try to classify specific tasks by the sorts of knowledge they require. A faculty approach to the mind has the advantage that we can work out general principles of its operation before considering the specific algorithms it exploits in individual cases.

Despite my optimism about top-down research I have also argued that a faculty approach to the mind is likely to be of little use in exploring central cognitive capacities. The study of central problems -- belief management, desire management, planning and the rest -- may not be susceptible to precise theorizing at the knowledge level. There is a note of irony in this. The central processes are mental processes which work primarily with occurrent knowledge. If someone were skeptical of the reality of knowledge, it is probable that he could forgive people thinking occurrent knowledge is real, but he would have little tolerance for those thinking tacit knowledge is real. If I am right, however, it is tacit knowledge, which, in the end, is the way into principled cognitive science. Where we have weak grounds for attributing tacit knowledge to a system, preferring instead to see the primary direction of processing to depend on interactions between fields of explicit and implicit belief, we have weak grounds for expecting knowledge level research to teach us much about mechanisms. This is not to say that knowledge level research into central processes is pointless. It is vital that we discover as much as possible about the beliefs and desires constituting the background conditions of thought etc. But knowledge level research can be expected to teach us very little about the management of these fields.

Different organisms may have different styles of management. Indeed I expect that attitude fields are regulated primarily by non-computational processes. Such was the view I have been entertaining these last two chapters. But all the facts have yet to come in. Further research is needed at the knowledge level to decide whether several systems of deduction, induction and abduction might not be sufficient to accomodate the laws of human thought.

These points, if true, are of some significance for a science of cognition still in search of stable foundations. I doubt, however, that they will be found convincing by many practising cognitive scientists. We need successful Knowledge Level theories in the tradition of Marr and Chomsky before we can expect much change in the way research is conducted by the members of the cognitive science community. No such theories have been presented here. Nonetheless, I hope enough has been said to vindicate those who might wish to try.

BIBLIOGRAPHY

Bibliographical information is provided here for all works quoted or referred to in the text, and to publications I have found particularly helpful. The dates given are those of the editions consulted, and do not necessarily show first publication.

Abelson, R.P. (1973)."The Structure of Belief Systems." In Schank, R.C. and Colby, K.M. (Eds.) **Computer Models of Thought and Language.** San Francisco: Freeman.

Allport, D.A. (1975)."The State of Cognitive Psychology; a Critical Notice of W.G. Chase. Visual Information Processing." **Quarterly Journal of Experimental Psychology** , 27, pp.141-52.

Anderson, J.R. (1976). **Language, Memory and Thought.** Hillsdale, NJ: Erlbaum.

Anderson, James and Michael Mozer (1981). "Categorization and Selective Neurons." in Hinton and Anderson, **Parallel Models of Associative Memory.**

Armstrong, D.M. (1968). A **Materialist Theory of the Mind.** London: Routledge & Kegan Paul.

Ayer, A.J. (1964). **Man as a Subject for Science.** University of London: Athlone Press.

Backus, J. (1978). "Can programming be liberated from the von Neumann style? A functional style and its algebra of programs." 21:613-641.

Bell, C.G. & Newell, A. (1971). **Computer Structures: Readings and Examples.** New York: McGraw-Hill.

Blanchard, Benjamin S. and Walter J. Fabrycky (1981). **Systems Engineering and Analysis.** Englewood Cliffs, NJ: Prentice Hall.

Block, N.J. (1980). **Readings in the Philosophy of Psychology** Vol 1. London: Methuen.

Block, N.J. (1981). **Readings in the Philosophy of Psychology** Vol 2. London: Methuen.

Bobrow, D.(1975)."Dimensions of Representation." in Bobrow, D. and Collins, A. (Eds.) **Representation and Understanding: Studies in Cognitive Science.** New York: Academic.

Boden, M.A. (1970)."Intentionality and Physical Systems." **Philosophy of Science.** 37, pp.200-14.

Boden, M.A. (1977). **Artificial Intelligence and Natural Man.** Hassocks: Harvester Press.

Bohm, David (1981). **Wholeness and the Implicate Order.** London: Routledge & Kegan Paul.

Brady, M. (1979). "Expert Problem Solvers." in Donald Michie (Ed.), **Expert Systems in the Micro Electronic Age.** Edinburgh University Press.

Braine, Martin D.S. (1978). "On the Relation Between the Natural Logic of Reasoning and Standard Logic." in **Psychological Review.** Vol.85, No.1, January 1978, pp. 1-21.

Braithwaite, R.B. (1953). **Scientific Explanation.** Cambridge: Cambridge University Press.

Broadbent, D.E. (1961). **Behaviour.** London: Eyre & Spottiswoode.

Broadbent, D.E. (1971). **Cognitive Psychology: Introduction.** British Medical Bulletin, 27, no.3, pp.191-4.

Broadbent, D.E. (1973). **In Defence of Empirical Psychology.** London: Methuen.

Campbell, D.T. (1950). "Common Fate, Similarity and other Indices of the Status of Aggregates of Person Social Entities." **Behavioural Science.** 14-25.

Campbell, D.T. & Fiske, D.W. (1959)."Convergent and Discriminant validation by the Multi-trait Multi-matrix Method." **Psychological Bulletin** 56, 81-105.

Chase, W.G. (Ed.) (1973) **Visual Information Processing.** New York: Academic Press.

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

Chomsky, N. (1959). "A review of B.F. Skinner's Verbal Behaviour." in **Language.** 1959, 35, pp. 26-58.

Chomsky, N. (1965). **Aspects of the Theory of Syntax.** Cambridge, Mass. MIT Press.

Chomsky, N. (1968). **Language and Mind.** New York: Harcourt, Brace, and World.

Chomsky, N. (1972). **Problems of Knowledge and Freedom.** Bungay, Suffolk: Fontana.

Chomsky, N. (1975). "Knowledge of language." in Gunderson, K. (Ed.) **Language, Mind and Knowledge.** Minneapolis, Minn. University of Minnesota Press.

Chomsky, N. (1980). **Rules and Representations.** Oxford: Basil Blackwell.

Churchland,P.M. (1979). **Scientific realism and the plasticity of mind.** New York: Cambridge Univ. Press.

Churchland, Patricia Smith (1980). "Neuroscience and psychology: should the labor be divided?" **Behavioural and Brain Sciences.** 3, p.133.

Colby, K.M. (1978). "Mind models: an overview of current work." **Mathematical Biosciences.** 39:159-185.

Craik, K.J.W. (1943). **The Nature of Explanation.** Cambridge: Cambridge University Press.

Davidson, Donald (1963."Actions, reasons, and causes." reprinted in **Actions and Events.**

Davidson, Donald (1966). "The Logical Form of Action Sentences." reprinted in **Actions and Events.**

Davidson, Donald (1970). "Mental Events." reprinted in **Actions and Events.**

Davidson, Donald (1971). "The Material Mind." reprinted in **Actions and Events.**

Davidson, Donald (1974). "Psychology as Philosophy." reprinted in **Actions and Events.**

Davidson, Donald (1980). **Essays on Actions and Events.** Oxford: Clarendon Press.

Davidson, Donald and Jaakko Hintikka (1969). (Eds.) **Words and Objections: Essays on the Work of W.V. Quine.** Dordrecht: D. Reidel.

Davis Martin (1978)."What is a Computation?", in Lynn Arthur Steen (Ed.) **Mathematics Today: Twelve Informal Essays.** New York: Vintage Books.

Dennet, Daniel C. (1978). **Brainstorms: Philosophical Essays on Mind and Psychology.** Montgomery: Bradford Books.

Dreyfus, H.L. (1979). **What computers can't do.** N.Y.: Harper & Row.

Dummett, Michael (1975)."What is a Theory of Meaning?" (I), in Samuel Guttenplan, (Ed.) **Mind and Language** , pp.97-138.

Dummett, Michael (1976)."What is a Theory of Meaning?" (II), in Gareth Evans and John Mcdowell, (Eds.) **Truth and Meaning.** pp.67-137.

Dummett, Michael (1978. **Truth and Other Enigmas.** Duckworth, London.

Ellis, Brian (1979). **Rational Belief Systems.** Basil Blackwell, Oxford.

Elster, Jon (1979). **Studies in Rationality and Irrationality.** Cambridge: Cambridge University Press.

Erickson, J.R. (1978). "Models of Formal Reasoning." in R.Revlin and R.E.Mayer (Eds.), **Human Reasoning.** Washington D.C.: Winston.

Evans, Gareth (1973)."The Causal Theory of Names." **Aristotelian Society Supplementary Vol.** xlvii, 187-208.

Evans, Gareth (1976). "Semantic Structures and Logical Form." in Gareth Evans & John Mcdowell (1976).(Eds.) **Truth and Meaning.** Oxford: Clarendon Press.

Evans, Gareth, & Mcdowell, John (1976).(Eds.) **Truth and Meaning.** Oxford: Clarendon Press.

Evans, Gareth, Varieties of Reference. Oxford : Clarendon Press (1982).

Fahlman, Scott (1981). "Representing Implicit Knowledge." in Hinton, Geoffrey and James Anderson (1981).(Eds.) **Parallel Models of Associative Memory.** Hillsdale NJ: Lawrence Erlbaum, 1981.

Farrell, B.A.(1950) "Experience", **Mind** , LIX.170-98.

Freeman, P.and A. Newell (1971). "A model for functional reasoning in design." **CMU-CS-71** -107.

Frege, Gottlob (1956)."The thought: a logical inquiry." reprinted in P.F.Strawson (Ed.) **Philosophical Logic.** London: Oxford Univ. Press, 1967.

Feigenbaum, E.A. (1979)."Themes and Case Studies of Knowledge Engineering." in Donald Michie, (Ed.), **Expert Systems in the Micro Electronic Age.** Edinburgh University Press.

Feynman, R. (1965). **The Character of Physical Law.** Cambridge M.I.T. Press.

Field, H.H. (1978). "Mental Representations." **Erkenntnis,** 13, pp.9-61.

Field, Hartley (1972). "Tarski's Theory of Truth." Reprinted in Mark Platts (Ed.) **Reference, Truth and Reality: Essays on the Philosophy of Language.** London: Routledge & Kegan Paul.

Fodor, Jerry (1975). **The Language of Thought.** Cambridge: Harvard University Press.

Fodor, Jerry A. (1980). "Methodological Solipsism Considered as a Research Strategy for Cognitive Psychology." **Behavioural and Brain Sciences,** 3, 63-73.

Fodor, J.A. (1981). **Representations.** Montgomery, Vermont: Bradford Books.

Fodor, Jerry A. (1965). "Functional Explanation in Psychology." **Readings in the Philosophy of the Social Sciences.** (Ed.) May Brodbeck.

Gauld, A.O. (1972). "The domain of Psychology." **Bulletin of the British Psychological Society,** 25, pp.93-100.

Gauld, Alan, and John Shotter (1977). **Human Action and its Psychological Investigation.** London: Routledge & Kegan Paul.

Geach, Peter (1957). **Mental Acts: Their Contents and their Objects.** London: Routledge & Kegan Paul.

Goffman, E. (1956). **The Presentation of Self in Everyday Life.** Edinburgh University Press. Harmondsworth: Penguin Books(1971).

Goldman, A.I. (1970). **A Theory of Human Action.** New York: Prentice Hall.

Goodman, Nelson (1965). **Fact, Fiction and Forecast.** Indianapolis: Bobbs-Merrill.

Greene, Judith (1972). **Psycholinguistics: Chomsky and Psychology.** Harmondsworth: Penguin Books.

Grossberg, S.(1980) "How does the brain build its code? **Psychological Review** ,:87-51

Guttenplan, Samuel (1975). **Mind and Language.** Oxford: Clarendon Press.

Haack, Susan (1978). **Philosophy of Logic.** Cambridge: Cambridge University Press.

Harman, Gilbert (1973). **Thought.** Princeton University Press, Princeton.

Harman, Gilbert H. (1977). "How to Use Propositions." **American Philosophical Quarterly.** xiv, 173-6.

Harré, R. & Secord, P.F. (1972). **The Explanation of Social Behaviour.** Oxford: Blackwell.

Harré, Rom and Mario Von Cranach (1982). (Eds.) **The Analysis of Action. Recent Theoretical and Empirical Advances.** Cambridge: Cambridge University Press.

Haugeland, J. (1978) "The nature and plausibility of cognitivism." in **The Behavioral and Brain Sciences.** 2, 215-260.

Haugeland, John (1980)."Programs, causal powers, and intentionality." in **The Behavioral and Brain Sciences.** (1980),3, pp.432-3.

Hayes, Patrick J. (1973). "The Frame Problem and Related Problems in Artificial Intelligence." reprinted in Weber and Nilsson (1981) **Readings in Artificial Intelligence.** Palo Alto: Tioga.

Hayes, P.J. (1979) "The Naive Physics Manifesto." in D. Michie (Ed.), **Expert Systems in the Micro Electronic Age.** Edinburgh University Press.

Heidegger, Martin (1962). **Being and Time.** translated by John Macquarrie and Edward Robinson, Oxford: Blackwell.

Hempel, Carl G. (1959). "The Logic of Functional Analysis", reprinted in **Aspects of Scientific Explanation.** pp.297-330.

Hempel, Carl G. (1965). **Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.** New York: The Free Press.

Higgenbotham, James (1982). "Noam Chomsky's Linguistic Theory." in **Social Research.** Spring 1982, Vol.49, No1. pp.143-157.

Hinde, Robert A. (1982). **Ethology.** Glasgow: Fontana Paperbacks.

Hinton, Geoffrey (1981). "Implementing Semantic Networks in Parallel Hardware." in Hinton and Anderson, **Parallel Models of Associative Memory.** 1981.

Hinton, Geoffrey and James Anderson (1981).(Eds.) **Parallel Models of Associative Memory.** Hillsdale NJ: Lawrence Erlbaum, 1981.

Hinton, Geoffrey and Terrence J.Sejnowski (1983). "Optimal Perceptual Inference." to appear in **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.**

Hoare, C.A.R. (1982). "Specifications, Programs and Implementations." **Technical Monograph PRG-29.**

Hook, Sydney (1960). **Dimensions of Mind.** New York: New York University Press.

Hubel, D.H. (1979). "The brain." **Scientific American.** 241(3),pp.44-53.

Kalke, William (Feb. 1969). "What is wrong with functionalism." **Noûs.** III, 1, 83-94.

Kemeny, John G., and Paul Oppenheim (1967). "On reduction." in **Readings in the Philosophy of Science.** (Ed.) Baruch A.Brody.

Kosslyn, S.M. (1981). "Imagery, Propositions and the Form of Internal Representations." reprinted in Ned Block(Ed.) **Readings in Philosophy of Psychology.** Vol.2, London: Methuen. 1981.

Levins, Richard (1966). "The Strategy of Model Building in Population Biology." **American Scientist.** 54(4):421-431.

Lewis, David (6 Jan.1966). "An argument for the identity theory." in **Journal of Philosophy.** LXIII, 1. 17-25.

Luria, A.R. (1969). "Human brain and psychological processes." **Mood states and Mind.** (Ed.) Karl H. Pribam.

Luria, A.R. (1973). **The Working Brain.** Harmondsworth, Middx. Penguin Books.

McCarthy, J. (1979). "Ascribing mental qualities to machines." In Ringle, M.(Ed.), **Philosophical Perspective in Artificial Intelligence.** Hassocks: Harvester Press.

McCarthy, J. (1979). "First Order Theories of Individual Concepts and Propositions." in D. Michie (Ed.) **Expert System in the Micro Electronic Age.**

McCulloch, W.S. and W.H. Pitts (1943). "A logical calculus of the ideas immanent in nervous activity." **Bulletin of Mathematical Biophysics.** 1943,5, pp.115-133.

McDowell, John (1977)."On the Sense and Reference of a Proper Name." **Mind.** lxxxvi, 159-85.

Mackie, J. L. (1974). **The Cement of the Universe.** Oxford: Oxford University Press.

Malcolm, N. (1968). "The Conceivability of Mechanism." **Philosophical Review.** 76, pp.45-72.

Marr, D. (1977). "Artificial intelligence-a personal view." **Artificial Intelligence.** 9:37-48.

Marr, David (1982). **Vision: A Computational Investigation into the Representation and Processing of Visual Information.** San Francisco: W.H.Freeman & Co.

Marr, D. (1976). "Early processing of visual information." **Phil.Trans.R.Soc.London** B275, 483-524.

Marr, D. (1982). "Visual information processing: the structure and creation of "visual representations." **Phil.Trans.R.Soc.London** B290, 199-218.

Marr, D. and H.K.Nishihara (1978). "Representation and recognition of the spatial organization of three-dimensional shapes." **Proc.R.Soc.London** B200, 269-294.

Marr D. and T.Poggio (1977). "From understanding computation to understanding neural circuitry." **Neurosciences Res.Prog.Bull.** 15, 470-488.

Marr D., T.Poggio (1979). "A computational theory of human stereo vision." **Proc.R.Soc.Lond.** B 204, 301-328.

Miller G. (1978). "Practical and Lexical Knowledge" in F.Rosch and B.B.Lloyd (Eds.) **Cognition and Categorization.** Hillsdale, N.J.: Lawrence Erlbaum, 1978.

Minsky, M.L. (1965). "Matter, Mind and Models." in **Proceedings of the International Federation of Information Processing Congress.** I, Washington, D.C., Spartan.

Minsky, Marvin (1975). "A Framework for Representing Knowledge." in Winston (Ed.). **Psychology of Computer Vision.** New York: McGraw-Hill.

Mischel, T. (1974). (Ed.) **Understanding Other Persons.** Oxford: Basil Blackwell.

Nagel, Ernest (1956) "A Formalization of functionalism." **Logic without Metaphysics.** By Ernest Nagel. Glencoe, Illinois: The Free Press.

Nagel, E. (1961) **The Structure of Science.** New York: Harcourt, Brace & World.

Nagel, Ernest, and James R. Newman (1971). **Gödel's Proof.** London: Routledge and Kegan Paul.

Nagel, Thomas (1965) "Physicalism" in C.V. Borst (Ed.) **The Mind-Brain Identity Theory.** London: Macmillan. 1977 pp.214-230.

Nelson, R.J. (15 July 1976) "Mechanism, Functionalism, and the Identity Theory." in **Journal of Philosophy.** LXXIII, 13, 365-85.

Newell, A. (1969). "Heuristic Programming: Ill-structured problems", in Aronofsky, J. (Ed.) **Progress in Operations Research** III, New York: Wiley.

Newell, A. (1973). "You can't play 20 questions with nature and win." in Chase, W.G. (Ed.) **Visual Information Processing.** New York: Academic.

Newell, A. (1973). "Production Systems: Models of Control Structures." in Chase, W.G. (Ed.), **Visual Information Processing.** New York: Academic Press.

Newell, A.(1980) "Reasoning, problem solving and decision processes: The problem space as a fundamental category." In R.Nickerson (Ed.), **Attention and Performance VIII.** Hillsdale, NJ: Erlbaum.

Newell, A.(1980) "Physical symbol systems." **Cognitive Science.** 4, 135-183.

Newell, A.,(1981) "Review of Nils Nilsson, Principles of Artificial Intelligence." **Contemporary Psychology.** 26,50-51.

Newell, A. & Simon, H.A. (1972) **Human Problem Solving.** Englewood Cliffs: Prentice-Hall.

Newell, A.,(1981) "The Knowledge Level." AAAI Society Presidential Address.

Newell, A. & H.A. Simon (1976) "Computer Science as empirical enquiry: Symbols and search." **Communications of the ACM.** 19(3), 113-126.

Nickles, T.(12 Apr. 1973) "Two Concepts of Inter-Theoretic Reduction." **Journal of Philosophy.** LXX, 7, 181-201.

Nowell-Smith, P.H.(1967) "Concept." In P. Edwards (Ed.) **The Encyclopedia of Philosophy.** New York: Macmillan: London, Collier-Macmillan. Vol. 2, pp.177-80.

Oatley, K.G. (1977)."Inference navigation and cognitive maps". in Johnson-Laird and Wason (Eds.) **Thinking: Readings in Cognitive Science** , Cambridge: Cambridge University Press.

Peirce, C.S. (1868). "Some Consequences of Four Incapacities." reprinted in J. Buchler (ed.) **Philosophical Writings of Peirce.** New York: Dover, 1955, pp.230-31.

Piaget, Jean (1971). **Biology and Knowledge.** Chicago: University of Chicago Press.

Palmer,S.F. (1978) "Fundamental aspects of cognitive representation." In E.H.Rosch, and B.B.Lloyd (Eds.) **Cognition and Categorization.** Hillsdale, N.J.: Erlbaum.

Peacocke, Christopher (1979). **Holistic Explanation.** Clarendon Press, Oxford.

Pike, K.L. (1967). **Language in relation to a unified theory of the structure of human behaviour.** The Hague: Mouton.

Pribram, Karl H. (1969). **Mood, States and Mind. Brain Behaviour.** Vol.1. Harmondsworth, Middx.: Penguin Books.

Prior, A.N. (1971). **Objects of Thought.** (Eds.) P.T. Geach and A.J.P. Kenny, Clarendon Press, Oxford.

Putnam, Hilary (1975). **Mind, Language and Reality. Philosophical Papers** , Vol.II. Cambridge: Cambridge University Press.

Putnam, Hilary (1975). "The Mental Life of Some Machines." , **Mind, Language and Reality .**

Putnam, Hilary (1975). "Minds and Machines." **Mind, Language, Reality .**

Putnam, Hilary (1975). "The Nature of Mental States." **Mind, Language and Reality.**

Putnam, Hilary,(1975). "Robots: Machines or Artificially Created Life?", **Mind, Language and Reality.**

Putnam, Hilary (1975). "Brains and Behaviour." in **Mind, Language and Reality.**

Putnam, Hilary (1975)." The Meaning of Meaning." in Hilary Putnam, **Mind, Language and Reality.**

Putnam, Hilary (1978). **Meaning and Moral Sciences** , London: Routledge & Kegan Paul.

Pylyshyn Zenan (1973). "What the mind's eye tells the mind's brain: a critique of mental imagery." **Physiological Bulletin.**

Pylyshyn, Z. (1978a) "Computational models and empirical constraints." in **Behavioural and Brain Sciences** 1, pp.93-99.

Pylyshyn, Z. (1978c) "Imagery and artificial intelligence." in W.Savage(ed.) **Perception and cognition: issues in the foundations of psychology.** Minneapolis: Univ. of Minnesota Press.

Pylyshyn, Z. (1979c) "The imagery debate analogue media or tacit knowledge?" **Psychological Review.** (1980).

Pylyshyn, Z. (1980) "Computation and cognition: Issues in the foundation of cognitive science." in **Behavioural and Brain Sciences** , 3, pp.11-169.

Quine, W. V. (1951) "Two Dogmas of Empiricism." in **From a Logical Point of View.** Cambridge: Harvard University Press. 1953 pp.20-46.

Quine, W.V. (1969) **Ontological Relativity and Other Essays** New York: Columbia University Press.

Quine, W.V. (1969)"Epistemology Naturalized", in W.V. Quine, **Ontological Relativity and Other Essays.** pp.69-90.

Quine, W.V., Ullian, J. S. (1970) **The Web of Belief.** New York: Random House

Revlin R. and V.O.Leirer (1978). "The effect of personal biases on syllogistic reasoning: Rational decisions from personalized representations." in Revlin and Mayer, **Human Reasoning** op cit.

Richards, D.A.J. (1971) A **Theory of Reasons for Action.** Oxford: Oxford University Press.

Rorty, Richard,(1972)" **Functionalism, machines, and Incorrigibility** ", Journal of Philosophy, LXIX,8, 203-20.

Rorty, Richard,(1965) " **Mind-Body Identity, Privacy and Categories** ", The Mind-Brain Identity Theory. (Ed.) C.V. Borst.

Ryle, Gilbert (1949) **The Concept of Mind.** New York: Barnes & Noble.

Ryle, Gilbert (1954). **Dilemmas.** Cambridge: Cambridge Univ. Press.

Sampson, Geoffrey (1980). **Making Sense.** Oxford: Oxford University Press.

Schaffner, Kenneth F. (June 1967) "Approaches to Reduction." **Philosophy of Science.** XXXIV,2, pp.137-47.

Schank, R. & Ableson, R. (1977) **Scripts, Plans, Goals and Understanding.** Hillsdale, New Jersey: Lawrence Erlbaum.

Schutz, Alfred (1976). "The Social World and the Theory of Social Action", and reprinted in his **Collected Papers** Vol.II The Hague: Martinus Nijhoff, 1976.

Schutz, Alfred (1976)."The Dimensions of the Social World",reprinted in his **Collected Papers** Vol.II The Hague: Martinus Nijhoff, 1976.

Schutz, Alfred (1976)."The Problem of Rationality in the Social World",reprinted in his **Collected Papers** Vol.II The Hague: Martinus Nijhoff.

Schutz, Alfred (1977). "Concept and Theory Formation in the Social Sciences" in Dallmayr and McCarthy, **Understanding and Social Inquiry.** Notre Dame: University of Notre Dame.

Searle, John R. (1969) **Speech Acts.** Cambridge: Cambridge University Press.

Shaw, R.E. (1971). "Cognition, Simulation and the Problem of Complexity". **Journal of Structural Learning** , 1971,2(4), pp.31-44.

Simon, H. (1969). **The Sciences of the Artificial** , Cambridge: MIT Press.

Simon, H. (1977). **Models of Discovery: and Other Topics in the Methods of Science,** Dordrecht: D.Reidel.

Simon, H. (1979). **Models of Thought,** New Haven: Yale Univ. Press.

Simon, H.A. (1980) "Cognitive Science: The newest of the artificial sciences." **Cognitive Science.** 33-46.

Sloman, A. (1979) **The Computer Revolution in Philosophy: Philosophy, Science and Models of Mind.** Hassocks: Harvester Press.

Smith, Brian Cantwell (1982). **The Computational Metaphor.** Unpublished draft; Cognitive and Instructional Sciences, Xerox Palo Alto Research Center.

Smith, Brian Cantwell (1982)."The Seven Percent Solution: An Analysis of the Central Problem in Artificial Intelligence and Some Suggestions about What to Do about It." Unpublished Draft.

Stalnaker, Robert (1976). "Propositions." in Alfred Mackay and Daniel Merrill (Eds.), **Issues in the Philosophy of Language.** (New Haven: Yale Press, 1976).

Sutherland, N.S. (1970) "Is the Brain a Physical System?" in R. Borger & F. Cioffi, (Eds.) **Explanation in the Behavioural Sciences.** Cambridge University Press.

Sufrin, Bernard (1981)."Formal System Specifications: Notations and Examples." in D. Neel. (Ed.) **Tools and Notions for Program Construction.**

Sufrin, Bernard (1981)."Formal Specification of a Display Editor." **Technical Monograph PRG- 21 .**

Tarski, Alfred (1969) "Truth and Proof" reprinted in Hanfling, O,(Ed.) **Fundamental Problems in Philosophy.** Oxford: Basil Blackwell.

Taylor, C. (1964) **The Explanation of Behaviour.** London: Routledge & Kegan Paul.

Taylor, Charles (1971). "Interpretation and the Sciences of Man." **Revue of Metaphysics,** 25, pp.3-51.

Taylor, Charles (1971) "Mind-Body Identity, a Side Issue?", in **The Mind-Brain Identity Theory.** (Ed.) C.V. Borst.

Thompson, D.W. (1917). **On Growth and Form.** Cambridge: Cambridge University Press.

Ullman, J.S. (1979). **The Interpretation of Visual Motion.** Cambridge,Mass.:MIT Press.

Waddington, C.H. (1972). **The Nature of Mind.** Edinburgh: Edinburgh University Press.

Waddington, C.H. (1973). **The Development of Mind.** Edinburgh: Edinburgh University Press.

Waddington, C.H. (1957). **Strategy of the Genes.** London: Allen and Unwin.

Wall, Robert (1972). **Introduction to Mathematical Linguistics.** Englewood Cliffs, NJ: Prentice-Hall.

Webber, Bonnie Lynn and Nils J. Nilsson (1981). (Eds.) **Readings in Artificial Intelligence.** Palo Alto: Tioga.

Weber Max (1977). "Objectivity in Social Science and Social Policy" reprinted in Fred Dallmayr and Thomas A. McCarthy, **Understanding and Social Inquiry,** Notre Dame: University of Notre Dame, 1977.

Weber Max (1977) "Basic Sociological Terms", reprinted in Fred Dallmayr and Thomas A. McCarthy, **Understanding and Social Inquiry,** Notre Dame: University of Notre Dame, 1977.

Weiskrantz, L. (1973) "Problems and Progress in Physiological Psychology", **British Journal of Psychology.** 64, pp.511-20.

White, A.R. (1968) **The Philosophy of Action.** Oxford University Press.

Wiggins, D. (1975)."Deliberation and Practical Reason." in Joseph Raz (Ed.) **Practical Reasoning** , Oxford: Oxford University Press.

Wimsatt, William C. (1976a) "Reductionism, Levels of Organisation and the Mind-Body Problem." in G.G.Globus, G. Maxell, and I. Savodnik, Eds. **Consciousness and the Brain.** New York: Plenum, 205-267.

Wimsatt, William C. (1974) "Complexity and Organisation." in K.F. Schaffner and R.S. Cohen, Eds. PSA-1972, **Boston Studies in the Philosophy of Science 20** (Dordrecht: Reidel) pp.67-86.

Winch, P. (1958) **The Idea of a Social Science and its Relations to Philosophy.** London, Routledge & Kegan Paul.

Winston, P.H. (1975). (Ed.) **Psychology of Computer Vision.** New York: McGraw-Hill.

Winston, P.H. (1975) "Learning structural descriptions from examples." in Winston, P.H. (Ed.), **The Psychology of Computer Vision.** New York: McGraw-Hill.

Young, R.M. (1979) "Production Systems for Modelling Human Cognition." in Donald Michie (Ed.), **Expert Systems in the Micro Electronic Age.** Edinburgh University Press.