CHRISTOPHER CHERNIAK

RATIONALITY AND THE STRUCTURE OF HUMAN MEMORY*

ABSTRACT. A tacit and highly idealized model of the agent's memory is presupposed in philosophy. The main features of a more psychologically realistic duplex (or *n*-plex) model are sketched here. It is argued that an adequate understanding of the rationality of an agent's actions is not possible without a satisfactory theory of the agent's memory and of the trade-offs involved in management of the memory, particularly involving "compartmentalization" of the belief set. The discussion identifies some basic constraints on the organization of knowledge representations in general.

What sort of models of the agent's memory are presupposed in philosophical psychology and the theory of knowledge? What is a more adequate, in particular, a less idealized, model? In answering these questions, I shall argue first that one cannot even explain important ranges of actual human behavior without employing a more "psychologically realistic" model. And second, I shall argue that an adequate understanding of the rationality of an agent's actions is not possible without such a model. The view that practical features of how we actually think are relevant to distinctively philosophical questions is by now a familiar one; examples are Quine's program for the "naturalization" of epistemology and Alvin Goldman's more recent proposal of a philosophical discipline of "epistemics".[1] And the natural initial response to this type of proposal is, I think, clear: Human beings are, no doubt, forgetful, careless, and so on; how can these failings be of interest to *philosophy*, as opposed to pedagogy, engineering, or other applied fields? The goal of this paper is to show that when we examine the structure of human memory in some detail, and when we employ a theory of *minimal*, as opposed to ideal, rationality, we do obtain conclusions which are directly germane to basic philosophical issues.

1. IDEALIZED AGENTS

Let us examine two examples of the type of idealization of the agent's memory which has been virtually universal in philosophy. The first example is Quine's influential model of the structure of the human belief system. That this idealization is present even within Quine's own

Synthese **57** (1983) 163–186. 0039–7857/83/0572–0163 $02.40

program for naturalizing the theory of knowledge illustrates the ubiquity of the idealization. In the last section of "Two Dogmas of Empiricism", Quine says, "total science is like a field of force whose boundary conditions are experience. A conflict with experience at the periphery occasions readjustments in the interior of the field". In particular, "reevaluation of some statements entails reevaluation of others, because of their logical interconnections . . .".[2]

A natural "descriptive" interpretation here is that Quine is claiming reevaluation of one belief entails reevaluation of others to maintain consistency of the total system, and that he is predicting the proprietor of a belief system will in fact make the appropriate reevaluations. Given Quine's naturalism, one would expect that his account is intended to be descriptively correct. This interpretation is confirmed by Quine's later discussion of the "interanimation of sentences" of a belief system in the first chapter of *Word and Object*.[3] Quine says of a portion of the belief system, "The theory as a whole – a chapter of chemistry, in this case, plus relevant adjuncts from logic and elsewhere – is a fabric of sentences variously associated to one another and to non-verbal stimuli by the mechanism of conditioned response". Again, Quine seems to be describing the interconnections among an agent's beliefs in terms of predicting changes in belief that really will happen, like the corresponding readjustments in the force field. Otherwise, the reference to "conditioned response" would not make sense.[4]

But the belief systems of actual human beings do not inevitably and automatically readjust themselves appropriately in the way Quine describes. The departures from Quine's idealization that we are concerned with here are certain types of forgetfulness; part of the human condition is in fact to fail to "make the connections" sometimes in a web of interconnected beliefs. For example, at least a decade before Fleming's discovery of penicillin, many microbiologists were aware that molds cause clear spots in bacteria cultures, and they knew such a bare spot indicates no bacterial growth. Yet they did not consider the possibility that molds release an antibacterial agent.[5] As we shall see, what makes this common kind of example philosophically significant, and not just an unfortunate case of human sloppiness, is how we can explain it. To begin with, we can say that the belief that molds cause bare spots seems to have been "filed" under the category of practical laboratory lore as information on undesirable contamination; the belief that a bare spot suggests inhibited bacterial growth seems to be in a

different file, on microbiological theory. Thus, the web of belief is not merely tangled; its fabric of sentences is "quilted" into a patchwork of relatively independent subsystems. Connections are less likely to be made between these subsets. The Quinian model does not take into account the basic organization of human memory.

As a second example of the idealization in philosophy of the memory structure of the agent, let us briefly consider the so-called "Preface Paradox".[6] If a person says F, "At least some of my beliefs are false", it seems highly likely that he will be correct. The conflict that is sometimes felt here arises because adding F to one's belief set guarantees that the belief set is inconsistent. The point that is of interest for us, and has been overlooked, is that the size of the belief set for which the person makes the statement of error F determines the reasonability of his joint assertions. If he says "Some sentence in $[p]$ is false, and p", this seems clearly irrational, like saying "I am inconsistent; I believe both p and not-p". If he says "Some sentence in $[p, q]$ is false, and p, and q", this is similarly unacceptable. But if the set is very large, and in particular is the person's total belief set, then accepting F along with that belief set becomes much more reasonable.

Why is the size of the belief set involved critical to whether the error assertion F is acceptable? If the person already thinks it is not merely possible, but very probable that some of his beliefs are inconsistent, then he loses little by adding F to his belief set. Now, what model of human cognition accounts for why a large belief set is very likely to be inconsistent? Among others, the model that began to emerge in the discussion of Quine's idealized web would explain this: The total belief set of a human being is so vast that he cannot even exhaustively enumerate its contents. Furthermore, it is organized into independent subsets; inconsistencies between elements in these different "files" are less likely to be detected. Thus, much of the perceived paradoxicality of the Preface Paradox seems to arise from presupposing an idealization of the agent's psychology that is very like Quine's; and, to that extent, that apparent paradoxicality can be dissipated by adopting a more adequate model of human memory.

2. A MODEL OF HUMAN MEMORY

Let us examine such a model. One model that is significantly more satisfactory than the idealizations prevalent in philosophy has been

employed in traditional psychology of verbal learning and memory for at least a century, continues to be fundamental in the more recent "constructive memory" and "semantic memory" approaches, and seems to be embedded even in our prescientific common sense explanations of behavior. The model does not apply transcendentally to all rational agents; it is easy to imagine agents that do not conform to the model, e.g., Quine's idealized agent. An attempt to abstract from the sometimes conflicting versions of these empirical theories must be oversimplified and incomplete. However, the resulting picture, even if crude, is an important improvement over the philosophical theory; we are concerned here with a matter of degree of idealization of theory. The standard model in the verbal learning and memory tradition[7] has a duplex structure; within a human's memory at any given moment a "short-term" or active memory and a "long-term" memory can be distinguished. One major aspect of this distinction reflects the everyday observation that a person cannot, at one moment, think about all the information he possesses; he can only consider a subset of it. The contents of the short-term memory correspond to what he is now thinking about, not necessarily consciously (as when I drive a car properly while conversing about something else); all other remembered information is in the long-term memory.

The storage capacity of short-term memory is commonly regarded as about six meaningful units or "chunks", such as randomly chosen words.[8] The duration of short-term storage of an item is also limited; the item is supposed to be remembered for less than half a minute if it is not "rehearsed" or repeated. The short-term memory is conceived of as a working memory, not just a passive store. That is, unlike long-term memory, operations can be performed on its contents, such as making deductive inferences from the activated beliefs there; in particular, the practical reasoning from beliefs and desires which results in undertaking an action can only occur there. Current constructive and semantic theories of memory include a similar conception of a short-term memory. It seems likely that there are several distinct special-purpose working memories and long-term stores. For example, a nonconscious working memory is a basic element of current psycholinguistic models of language comprehension and production. The "duplex" model can therefore be generalized as "n-plex", where $n > 1$.

In contrast to the span of short-term memory, the long-term memory is generally regarded as having no practical capacity limit; also, some

information can be stored there indefinitely. Therefore, long-term memory is where the vast majority of a person's "belief system" is at any given time. Items can be recalled or retrieved from long-term memory to short-term memory, that is, copied into short-term memory without being erased from long-term memory. According to the traditional models, items in long-term memory are in "cold storage" and are virtually dormant; they can undergo none of the processing that items in short-term memory can (the model can be modified to include degrees of activation, as opposed to "all or nothing"). In particular, items there cannot affect behavior; for a belief to influence actions, it must first pass through the great bottleneck of short-term memory.

All of the accounts of human memory we are considering claim that the contents of long-term memory are organized. An item in long-term memory is located for retrieval not by a search of the entire memory, but by a narrower search that takes advantage of the structure of the memory. All of these accounts in effect represent the long-term memory as a graph-theoretic entity, a network of nodes interconnected by arcs. The model is a generalization of the notion of a filing system, where a file can in turn contain subfiles. Each node is a storage location, containing a bundle of information. A search of long-term memory proceeds from node to node, via the interconnecting pathways; the search can be compared to running a maze. In the pure traditional theory, the interconnections are associative links only; a familiar example from Hume is that, given one idea, which other ideas come to mind may be determined by bonds formed by past experience of the conjunction of that idea with other ideas. Consequently, much memory organization may be idiosyncratic to the particular person.

More recent "semantic memory" accounts,[9] which have been developed in computer modeling of human memory, include so similar a picture of the structure of long-term memory that they are sometimes characterized as "neoassociationist". They generalize the interconnecting arcs to represent different kinds of relationships among the nodes; for instance, a directed arc between nodes n and m may represent that the item at n is an instance of, as opposed to a property of, the item at m. "Constructive memory" accounts[10] emphasize that what is stored in long-term memory is not a fixed replica of an experienced event, e.g., the exact wording of an encountered sentence. Instead, the underlying meaning of the sentence is integrated into the current memory representations or "schemata". These are therefore more than just a filing system

for specific memories (the model resembles in some respects the Kantian account of perception). Recall similarly is supposed to be a synthesizing process, in which one reconstructs, in the working memory, "what must have occurred" from a few fragments stored in long-term memory. For our purposes, the most important divergence of the constructive approach from traditional accounts is that for the former, information in long-term memory is not inactive; after acquisition and before recall, a proposition may be assimilated to previous knowledge. The point remains that, for the constructive approach, processing in long-term memory is much more limited than processing in short-term memory (for some constructive accounts, this transformation of the stored item in fact occurs in a limited capacity working memory).

3. COMMON-SENSE PSYCHOLOGY

What model of human memory is presupposed by our common sense intentional explanations of behavior? I have argued in 'Minimal Rationality' that we are able to understand and predict quite successfully an agent's actions on the basis of a prescientific cognitive theory that attributes a system of beliefs, desires, and other intentional states to him. The very possibility of a predictive cognitive science traditionally seems to have been denied because of covert acceptance of a conception of rationality so idealized that it is, for most purposes, not at all applicable to actual human beings. Our tacit everyday cognitive theory must include some other type of rationality condition; a concept of *minimal* rationality, where the agent can have a less than perfect ability to choose appropriate actions, is needed. Such an ability to identify appropriate actions in turn implies possession of some logical insight: in particular, an ability to make some, but not all, inferences from one's beliefs that are useful for this purpose (satisfaction of a minimal inference condition); and an ability to eliminate some, but not all, inconsistencies that arise in one's belief set (satisfaction of a minimal consistency condition).

A first approximation of the prediction scheme involved here is: the observer has attributed a particular set of beliefs and desires to a putative agent. Just to qualify as having that belief-desire set, the agent must attempt some of the actions which, according to those beliefs, would tend to satisfy those desires. That is, the agent must act minimally, but not necessarily ideally, rationally. The observer can identify this required set of actions, and if he has attributed the correct belief-desire set to the

agent, predict the agent's actions. However, this sketch of the predictive schema is still very incomplete. To predict an agent's actions in any interesting detail, the observer must know not just that the agent will undertake some of the apparently useful actions; the observer must be able to some extent to determine *which* ones. In fact, the minimal rationality conditions in everyday practice are embedded in a broad range of other cognitive psychological theories which fill in where the agent's behavior will depart from ideal rationality. One of the most important of these is a theory of human memory structure. (Another is a psychological theory of the relative difficulty of different inferences for the agent.)[11]

While this common-sense theory of human memory may be more primitive than the theories of present-day "scientific" psychology, it seems to share the two main elements that we have found in the latter theories. First, the tacit common-sense theory includes a short-term/long-term memory distinction: Only a small subset of one's total belief system, as the contents of short-term memory, can be activated or thought about at a given time; only these can influence the choice of actions, and in particular, be "logically processed" – used as premises for inferences or compared for inconsistency. We can only "make the connections" for these. The rest of the beliefs, those in the long-term memory, are relatively inert. Second, the common-sense theory assumes an organization of long-term memory, one that determines the pattern of a search for an item, and leads to some failures of retrieval to short-term memory. To predict significantly the agent's behavior on the basis of an attributed belief-desire set, we need to know which beliefs (and other elements of his cognitive system) are now in short-term memory, since they will otherwise be inactive; we seem to do so by using this memory model.

We employ the memory model to understand some very prevalent lapses from ideal rationality. I will not attempt an exhaustive typology of such lapses. One important kind of example: Jones may have proven that $a = b$ and also that $b = c$ at different stages of a long derivation. If asked whether $a = b$, Jones would then assent, and similarly for $b = c$. Jones has not forgotten either sentence; he still believes both of them. And yet it is a common enough situation for Jones then to wonder whether $a = c$, and not to be able to find out. When this occurs in ordinary life, we do not feel a temptation to say Jones didn't believe that $a = b$ or didn't believe that $b = c$. For we can make sense of this failure to make a very obvious

useful inference from the two beliefs in terms of the short-term/long-term memory distinction: When Jones reached the point in the proof where he asked himself whether $a = c$, he was not then thinking about $a = b$ and $b = c$. It was not the case that both of these beliefs were in short-term memory; hence, one or more of them was inactive, not capable of being used as a premise in reasoning.

Failure to acknowledge the short-term/long-term memory distinction seems responsible for most of the common denials in philosophy of the possibility of people making obvious logical errors.[12] An important example, as one might expect from our discussion of Quine's model of consistency maintenance, is Quine's Principle of Charity, in particular, his thesis that correct interpretation of a person's utterances must not attribute inconsistencies to him. Similar issues are involved in debates about whether a person can believe obvious contradictions. Another important case, especially prevalent in philosophical analyses of decision and game theory, is the requirement of perfect preference transitivity; an agent supposedly cannot ever prefer a over b, b over c, and c over a. A further example is the frequent claim that it makes no sense to say a person believes p, believes $p \rightarrow q$, yet does not believe q.[13] All of these cases can be explained on the same pattern as the Jones example.

We also seem to explain some everyday behavior in terms of our theory of the specific organization of a given individual's long-term memory. For instance, Smith believes an open flame can ignite gasoline (he uses matches to light bonfires, etc.), and Smith believes the match he now holds has an open flame (he would not touch the tip, etc.), and Smith is not suicidal. Yet Smith decides to see whether a gasoline tank is empty by looking inside, while holding the match nearby for illumination. Similar stories often appear in newspapers; this is approximately how one of Faulkner's characters dies in *The Town*. The anecdote at the beginning of this paper about the non-discovery of penicillin, and Duncker's classic experiments on problem-solving[14], involve the same important type of lapse.

We can explain Smith's failure to infer the obvious conclusion that his match might ignite the gasoline by use of a very plausible hypothesis about the taxomony by which Smith's beliefs are organized. We seem to assume that in Smith's not especially idiosyncratic categorization scheme, the belief that a flame can ignite gasoline is filed under, roughly, "means of ignition"; the belief that the match he now holds has a flame

has been classified instead under "means of illumination". The "illumination" category rather than the "ignition" category was checked because Smith decided he needed more light to see into the tank. The two crucial beliefs here (along with others) therefore were not both in short-term memory to be "put together"; but only if they were being thought about together could Smith make the connection and infer that there was danger. In this way, some of a human agent's departures from perfect rationality follow predictable patterns, understandable in terms of the organization of his long-term memory.

Less than ideally rational behavior like Jones's and Smith's is an uncontestable major feature of actual careful science, e.g., many medical misdiagnoses, as well as of sloppy everyday life. However, instead of explaining, say, Smith's behavior in terms of the structure of his memory, one might try to argue that, at the time Smith lights the match, either he does not really believe that the match has a flame, or else he does not then believe that a flame can ignite gasoline. For instance in the former case, Smith might just think the match was an illumination source, and have no opinion about whether it was an ignition source.

One problem for this alternative account is that, if it is not to be just an *ad hoc* explanation for this one case only, it will entail a conception of a peculiarly indecisive agent. An agent often acts, as we would ordinarily say, inappropriately for one of his beliefs, whether because of forgetfulness, failure to infer a consequence of the belief, or for other reasons. Each time one of these inappropriate actions is followed by an action appropriate for the belief and vice versa, we would have to say that the agent had changed his mind regarding the proposition in question. Over an interval when we would normally claim the agent had one stable, enduring belief, we would instead have to say he very repetitiously kept temporarily changing his mind back and forth. It does not seem arbitrary here to prefer an explanation of the agent's behavior in terms of his memory structure to an account that makes the agent's behavior just a patternless coincidence of wavering. (Of course, this does not imply that there cannot be genuine changes of opinion.)

I think the main motivation for the view that, at the fatal moment, Smith has no opinion about whether his match is an ignition source, and for a similar treatment of the Jones case, is acceptance of an ideal rationality requirement: Smith can't believe the match has a flame because if he did, he must – by the ideal rationality condition that he make all useful inferences – conclude that holding it near the tank is

dangerous; and he doesn't do this. But, as mentioned earlier, such an idealization requires the agent to have unlimited cognitive resources (of memory and time), and so according to it, Smith cannot in fact have *any* beliefs. The ideal rationality conditions are unacceptable, and so they are not a satisfactory basis for rejecting our earlier explanation in terms of memory organization. Thus, one way in which the psychology of memory is philosophically relevant is now evident. Without something like the memory model we have been exploring, philosophical accounts cannot explain, or even admit the possibility of, a large and important range of human behavior, involving making obvious mistakes.

4. TWO STANDARDS OF RATIONALITY

Let us now turn from the descriptive adequacy of this model to the normative issues of how a memory ought to be organized and of which actions an agent ought to undertake. One consequence of the model of human memory structure implicit in our common-sense cognitive theory is that there are two distinct levels of minimal rationality, one required for a person's inactive belief set, and another more stringent one required for his current activated belief set. For the short-term/long-term memory distinction entails that only beliefs in short-term memory can be premises in reasoning; beliefs in long-term memory are inert – they don't interact with each other, and they don't affect behavior. Correspondingly, while beliefs in long-term memory are not free of all rationality constraints, more rationality is required of the beliefs in short-term memory. For example, we found no difficulty understanding how Jones could believe $a = b$ and believe $b = c$ without inferring $a = c$, so long as the two beliefs were not both activated at the same time. But this would not be true if Jones were then considering both of these beliefs. If we asked him whether he realized that he had proven that $a = b$ and that he had proven that $b = c$, and he still claimed he didn't see that $a = c$, we would conclude typically either that he did make the inference – perhaps his claim was not sincere – or else that the two beliefs had not in fact been activated – perhaps he did not understand our question. That is, we do not always require even the most obvious useful inference to be made from inactivated beliefs; but generally before we will accept that such an inference is not made from those beliefs when they are activated, we will reappraise the supposition that the beliefs are activated.

One might think the higher standard of rationality for short-term

memory is just an idiosyncrasy of our common-sense cognitive theory that, fortunately, reflects the psychological facts of how our minds happen to operate. For instance, W. J. McGuire, in a paper in the empirical psychology of beliefs and attitudes, presented a model and several experiments on how people maintain consistency in their cognitive systems.[15] McGuire sought empirical support for the claim that a person's belief set is subject to a "Socratic Effect", that is, "a person's beliefs on logically related propositions can be modified by the Socratic method of merely asking him to verbalize his beliefs, thereby sensitizing him to any inconsistencies among his beliefs, and thus inducing changes toward greater internal consistency" (p. 79). McGuire did not seem to be aware that much of his model is a special case of the conventional model of human memory structure we have outlined; the "arena of consciousness" into which beliefs are recalled by the "Socratic method" and where they are compared for consistency corresponds to the short-term working memory. The Socratic Effect is an example of the higher level of rationality that applies to the contents of the short-term memory.

But a distinction must be made between a mere empirical generalization and a fundamental element of our cognitive theory. A more obvious case is that McGuire and many other investigators in the psychology of attitudes and beliefs include in their models an assumption that a person has a "need for consistency" and will attempt to maintain consistency in his belief set; they then treat this assumption as a low-level empirical hypothesis requiring experimental confirmation. But, while there are valuable experimental questions in the field, the investigators seem unaware that if a putative agent does not attempt to maintain some consistency among his supposed beliefs, we will deny that he has any beliefs at all. As mentioned earlier, this is one of the minimal rationality conditions on any belief system; we can be sure that it applies before we try any experiments.

Similarly, it is not just a psychological accident that the human short-term memory is subject to more stringent rationality conditions than the long-term memory. For, given that beliefs in human long-term memory are virtually inactive, the only opportunity for the beliefs of the total system to be logically processed and kept rational at all is in what happens to the belief subset in short-term memory. Whatever may be the minimum "passing grade" of rationality for the contents of long-term memory in a given case, if the activated subset did not exceed it, the long-term memory could not be maintained at that minimum level. This

is because, as we have seen, the inactivity of beliefs in long-term memory in itself degrades rationality. It results in the accumulation of unrecognized inconsistencies, valuable inferences not being made, and so on. Only the behavior of the contents of short-term memory can counterbalance the results of the inertness of the beliefs in long-term memory, and so contribute to the maintenance of adequate rationality.

Thus, the Socratic Effect seems more than just a natural law of the human mind, such as the fact that our normal short-term memory capacity is six rather than twelve "chunks". Given the low quality of processing of the contents of human long-term memory, if someone were not more likely, e.g., to discover inconsistencies among some subset of his beliefs, he would not be able to maintain enough rationality to qualify as having beliefs. This conclusion should apply for any creature, of human psychology or otherwise, that has a similarly duplex memory.

5. EFFICIENT RECALL

A correspondingly general conclusion applies to the other main element of the human memory model, the organization of long-term memory into independently activated subsets. I will argue that it is not just an accident of human psychology that the long-term memory is so organized; given the short-term/long-term memory distinction and some other basic constraints, the long-term memory could not be otherwise, or the total belief system could not maintain minimal rationality.

The first step in connecting long-term memory organization to minimal rationality is to note that minimal rationality of a duplex cognitive system requires efficient recall of items from the inert long-term memory to short-term memory. If only beliefs in short-term memory can control decisions, then a person cannot act minimally rationally – that is, to some extent, appropriately according to his beliefs – unless, at least sometimes, the "right" beliefs are recalled to short-term memory. The right beliefs here are those that are relevant to making a current decision about whether or not to undertake a given action. For instance, it seems that just this type of failure to recall appropriate beliefs has occurred when Smith holds the match near the gasoline tank; given Smith's goal of self-preservation, his beliefs about whether a flame will ignite gasoline, etc. are obviously relevant to the question of whether or not he should attempt this action.

Of course, Smith can still qualify in this case as having a belief system

because his recall capacity, although not perfect in reliability and speed, might not be entirely inadequate. The most inadequate recall capacity would be a null one, where no putative "beliefs" were recalled from long-term to short-term memory; there would then in fact be no actual long-term memory. The next worst recall procedure would be one which was on a completely random basis; that is, where "beliefs" were recalled, but generally were unrelated to the current contents – e.g., goals and beliefs – in the short-term memory. Even if the contents of such a short-term memory (of normal human size, relative to long-term memory) satisfied some ideal rationality condition (for example, that any desired logical operation could always be performed), the relation of the putative total belief system to attempted actions would clearly disintegrate into chaos and fail to qualify as at all rational. Thus, there is some lower bound on recall capacity.

How can the recall efficiency required for the minimal rationality of a cognitive system be achieved? This would be relatively simple if a person's belief system could be exhaustively searched in each case. For instance, in the situation Quine describes in the passages cited earlier, if, whenever I decided to add a sentence p to my belief system, I could check the consistency of p with every subset of my current total belief set. Given unlimited time, we could in principle perhaps make such complete "algorithmic" searches to locate a desired item. But in fact, it is commonly assumed in the psychology of memory that we cannot retrieve in this way;[16] the simple fact that we fail to recall desired information alone suggests this. The storage capacity of human long-term memory has no well-defined upper bounds. The conventional view is that there are too many beliefs in the long-term memory for exhaustive searches. Further, the time available in which to identify desirable actions before the opportunity to benefit from them has passed is too limited. Even with perfect retention, for a super-mnemonist, the problem of locating a desired item would remain, and indeed be worse.

Nonetheless, the unfeasibility of exhaustive memory search is not as basic a feature of our psychology as, for example, our finite memory capacity. A cognitive system cannot be arbitrarily small; no one could just have the single belief that $2 + 2 = 4$. The holistic point is that a putative cognitive system must attain a certain "critical mass" before it can include any beliefs. However, this still does not exclude the possibility of cognitive systems that are simple enough, and have rapid enough search procedures, so that exhaustive search would be

practically possible. The point remains that the problems of maintaining rationality for such a creature would be fundamentally unlike those for a human being with a normally rich cognitive system. For instance, it would not be advisable for a normal human even to attempt an exhaustive search, if there were other valuable uses of his limited cognitive resources.

In the section of the *Treatise of Human Nature* on abstract ideas,[17] Hume recognizes the "given" for human beings of the unfeasibility of algorithmic memory searches. He says, "as the production of all the ideas to which [a] name may be applied, is in most cases impossible, we abridge that work by a more partial consideration, and find but few inconveniences to arise in our reasoning from that abridgement" (p. 21). The question now is, on what basis can the search be abridged, given that we know that a random narrowing would lethally "inconvenience" our reasoning? Hume is also aware of this problem: "Nothing is more admirable, than the readiness, with which the imagination suggests its ideas, and presents them at the very instant, in which they become necessary or useful" (p. 24). However, Hume seems to overlook the same kind of point we have found overlooked earlier, that this "most extraordinary" ability is more than just a fact of our particular psychology; it seems a precondition for us to qualify as having beliefs (or ideas or concepts). Nor does Hume attempt to explain what selection mechanism would be able to work so well; he just concludes that it is "a kind of magical faculty in the soul".

6. CHOOSING AN INQUIRY

Hume's problem has a wider philosophical significance. One may ask, how can I decide to recall to short-term memory a specific item currently in long-term memory, without *already* having the item available? The more general problem is deciding what to think about next. And this in turn is an instance of a rarely acknowledged but fundamental epistemological question, "what should I inquire about?" The answer must be relative to my current goals and beliefs about how to attain those goals. The most important point is that a creature that did no better than chance in identifying inquiries that were valuable for it would be a radically defective agent. Making a deductive inference was the kind of inquiry 'Minimal Rationality' focused upon, as opposed to undertaking an empirical investigation, and the conclusion there in fact applies here:

this creature would in particular be a heuristic imbecile in its deductions, and so could not be minimally rational – that is, could not even be an *agent*.

The question of what we should inquire about cannot be ignored, because we are each in the finitary predicament, with cognitive resources that are severely limited relative to the range of possible inquiries. We cannot obtain and use all available information. Furthermore, it would not be advisable to attempt to do so. Collecting much of this information would not be reasonable, because it is of no forseeable value at the time, and collecting it prevents the epistemic agent from using his limited resources for other activities which are obviously valuable only at that time. A person could waste his entire lifetime collecting only such uninteresting information. We must therefore try to determine the best use of our resources by deciding which information would be most useful to seek.

A predicament arises because the outcome of a decision about whether or not to obtain a parcel of information cannot be guaranteed in advance. Undertaking an inquiry is like undertaking any other action; it entails risks, costs, and benefits. The solution for this general problem, as well as for the special case of memory search, will involve doing better than chance, but not, of course, doing perfectly; "undershooting" and "overshooting" are unavoidable. For, the basis for deciding what inquiry to pursue will typically be incomplete. Otherwise, no decision about further inquiry may remain – the agent may end up already having made the inquiry. In addition, there is a point of diminishing returns, beyond which there are better uses of the agent's resources than in perfecting his choice of inquiry. And since such evaluations are themselves a particular kind of epistemic project, the selection of such projects cannot itself always be on the basis of an actual inquiry, or there will be a regress. (The regress seems in fact to be ended for human beings by their having been constructed, as a result of natural selection, so that certain inquiries are undertaken automatically and instinctively.)[18]

An adequate evaluation strategy must deal with this dilemma of having to be neither too restrictive nor too lax. In a passage in the *Critique of Pure Reason*, Kant seems to be aware of the general problem of determining the value of an inquiry, but not to perceive the dilemma involved.

> . . . in the endeavor to extend our knowledge a meddlesome curiousity is far less injurious

than the habit of always insisting before entering on any enquiries, upon antecedent proof of the utility of the enquiries – an absurd demand, since prior to completion of the enquiries we are not in a position to form the least conception of this utility, even if it were placed before our eyes.[19]

Kant regards any evaluation strategy as unfeasible, and he seems to imply that none is needed – that is, that the most liberal possible one would suffice, contrary to what we have found. In fact, some inquiry-selecting ability of this kind is a precondition for being an agent. And, we *are* able to "form the least conception" of the value of prospective inquiries – for instance, in deciding sometimes whether a given area of investigation is relevant to a current project. Given that an agent cannot have a "tunnel intelligence", how can the "corner of his mind's eye" enable him to make choices that are better than completely random gropes? Let us return to the problem of efficient recall as an instance of this predicament.

7. NONALGORITHMIC SEARCH

What is required for adequate recall is a satisfactory heuristic, as opposed to algorithmic, search strategy. Since long-term memory is unmanageably large for exhaustive search, the strategy should involve fully searching only a subset of the items in memory. It is a commonplace in the psychology of memory, as well as in the management of other large information systems, such as libraries and computer memories, that suitable organization of the stored items makes locating items relevant to a particular question easier. This is the underlying rationale for the network structure that we saw generally proposed in models of human long-term memory, as well as for the cataloguing systems of libraries. In computer science, in particular in artificial intelligence, the variety of search schemes is vast; the importance of "the problem of knowledge representation" and, in particular, the need to subdivide large knowledge representations has long been recognized.[20]

As in the case of the general problem of choosing an inquiry, the required strategy here must be better than chance, but need not, of course, be perfect; the latter would require prescience. Searches can be expected to fail frequently in either possible way: beliefs that turn out not to be currently relevant may be checked, and beliefs that turn out to be useful may be skipped. How can this strategy have a better chance of success than random search – a search of an arbitrarily selected subset?

For this purpose, the stored items must be organized into subsets according to subject matter, where items within a subset are more likely to be relevant to each other than items from different subsets.

Which of a given set of items should be stored and grouped together in this way, that is, which are related to each other, is not an objective matter; it need not be the same for every rational creature. To some extent, how the items should be organized depends on which kinds of searches are most often made, which in turn will depend on the questions the creature is likely to ask, because of its beliefs and goals. For instance, if p and q together imply r and it is useful, given the agent's desires and beliefs, for the agent to find out that r is a consequence of his beliefs p and q, then to that extent, it is advisable for p and q to be in the same subset; however, this would not be so if the logical relationship among p, q, and r was not of interest to the agent. On the other hand, the inconsistency of a set of beliefs always makes them, to that extent, "objectively" relevant to each other, whatever the agent's other beliefs and desires, because of the rationality requirement that minimal consistency be maintained. For human beings, some of the basic features of this structuring should be the result of natural selection, since they would have been helpful for survival in any likely terrestrial environment (although they may cease to be, as the individual departs from hunter-gatherer conditions). The rest of the individual's particular organizational scheme is learned, some of it as part of one's culture, but much of it as idiosyncratic and flexible cognitive habits based on past experience.

Unlike the algorithmic strategy, this strategy cannot be guaranteed to succeed, but it would be faster. It is a tradeoff of reliability for speed, one of a series of tradeoffs of competing desiderata involved in satisfactory organization of memory. As we shall see, another "golden mean" concerns the size of the subsets. If they are too large, then exhaustive searches within a subset will take too long and will approximate the algorithmic strategy for the entire system; if they are too small, then the chance of selecting the wrong subset, and so missing a desired item, will be too great. (This dilemma is ameliorated somewhat, but not eliminated, by hierarchically nesting the subsets.)

In terms of the rationality conditions on a cognitive system, the major cost of structuring the contents of long-term memory in this way is that inconsistencies and useful inferences that involve beliefs in different subsets are likely to be unrecognized. For example, McGuire points out, "The appearance and persistence of cognitive inconsistency in the

individual indicate a degree of 'logic-tight' compartmentalization in the human thinking apparatus, by virtue of which certain sets of cognitions can be maintained isolated from one another, without regard for their logical interrelatedness" (p. 98). Logical relations between beliefs in different "compartments" are less likely to be recognized than relations among beliefs within one compartment, because in the former case the relevant beliefs are less likely to be contemporaneously activated, and, as we have seen, it is only when they are activated together that such relations can be determined. The result is that, as Herbert Simon had noted much earlier in another connection, actual human behavior "exhibits a mosaic character", a patterned lack of integration; "behavior reveals 'segments' of rationality...behavior shows rational organization within each segment, but the segments themselves have no very strong interconnections".[21]

As in his discussion of the Socratic Effect, McGuire does not seem to appreciate the fundamental status of compartmentalization. He refers to the well-known studies that suggest "authoritarian" types of personality favor compartmentalization as an ego-defensive strategy, and proposes that the "cognitive barriers" between compartments can be made "more permeable" by the Socratic method of asking subjects to state inconsistent opinions in close temporal contiguity. However, McGuire's compartmentalization of beliefs is just a specific instance of the general organization of items in long-term memory into subsets. And we have seen that some degree of such structuring seems to be an indispensable feature of a satisfactory heuristic search strategy, rather than an easily eliminable human flaw. We can now appreciate both the costs and the benefits of this strategy; prima facie, the resulting behavior can be characterized as departures from rationality, but on the assumption that exhaustive memory search is not feasible, such memory organization is advisable overall, in the long run, despite its costs. Correspondingly, a person's action may seem irrational when considered in isolation, but it may be rational when it is more broadly considered as part of the worthwhile price of good memory management.

8. DIMINISHING RETURNS

The search strategy of structuring the memory as subsets of related beliefs is a matter of degree. Two items are in different subsets or compartments if they tend not to be recalled together. There are then at

least two ways in which one belief system may be more compartmentalized than another. First, one of the systems might have more compartments than the other. Second, both systems might have the same pattern of compartments, but in one, some of the compartments might be less "permeable"; the compartmentalization would be greater in that beliefs from different compartments would be less likely to be recalled and considered together. We now know that (I) if there is no compartmentalization, if there is an equal likelihood that any belief will be recalled in conjunction with any other belief, cognitive resources will be spread too thin. Some degree of compartmentalization is indispensable for adequate management of our large memories; otherwise recall would be too poor for the supposed cognitive system of which the memory is a part to satisfy the rationality conditions.

But we have also seen that the cost even of useful compartmentalization is in unreliability of recall. And it is now clear that too much compartmentalization will be counterproductive for efficient recall. Extreme compartmentalization can exclude a belief system from satisfying the rationality conditions in two different ways. On the one hand, (II) if a would-be belief system is organized into too many "sharply defined" small compartments, it in effect disintegrates into unrelated fragments. Too many of the "beliefs" will be unlikely to be activated together, with the result that too many inconsistencies and useful inferences involving them cannot be recognized.

On the other hand, (III) if much of a putative belief system is organized into a few sharply defined large compartments, we may feel that, instead of chaos, there is a "split personality" with corresponding total belief sets that are each employed in different types of situations. These sets can overlap considerably and still represent distinct persons, if the inconsistencies and missed inferences within each of these sets are sufficiently fewer than those in the conjoint total set. It is commonly pointed out in the philosophy of mind that the set of mental entities that constitutes a person must fit together in a particular type of coherent whole. The point here is that a special case of this required integration is that a person's beliefs must satisfy the rationality conditions; in this way, too much compartmentalization of a cognitive system violates our concept of a person.

Thus, as compartmentalization increases, there is a kind of diminishing return. Not only is there a minimum limit on compartmentalization for adequate search efficiency, but there is also a maximum limit. The cost of

compartmentalization is some isolation of subsets of the belief system from each other, and the resulting lack of interaction can fragment the total system. The contents of long-term memory are subject to less stringent rationality requirements than the contents of short-term memory, but they are not permitted unlimited irrationality. Only a balance of compartmentalization of long-term memory enables a complete cognitive system to qualify as minimally rational. Given the small capacity of the short-term memory in which all higher quality processing must occur, and the unfeasibility of exhaustive search of long-term memory, "moderate" compartmentalization is required for any rationality.

It is also clear now that (IV) a candidate for a belief set can fail to be adequately rational in another way, if it is just compartmentalized in "the wrong way", rather than too much or too little. Two sets can be compartmentalized equally in the above sense, and yet one may be adequately rational, while the other is not. The latter set would fail to be rational because it was not organized into subsets of related beliefs, that is, it was not so organized as to satisfy inference and consistency conditions: too many apparently useful inferences and too many inconsistencies were not recognizable. This in turn would be because the sets of supposed beliefs involved in those inferences and inconsistencies were not grouped together, and so were unlikely to be contemporaneously recalled. The fact that the "right" way to organize a cognitive system to some extent depends on the individual's desires and beliefs does not imply that any organizational scheme at all will be equally adequate for that individual.

We now have the solution to Hume's mystery of how nonalgorithmic memory search procedures can be adequate; no magic homunculus is necessary. We have found a connection between memory organization and rationality: a basic precondition for our minimal rationality is (a) efficient recall, which itself requires (b) nonalgorithmic search, and that in turn requires (c) compartmentalization. Compartmentalization is in this way a fundamental constraint on human knowledge representations. Even if there might be other satisfactory ways of solving Hume's problem, compartmentalization is not just a regrettable failing of human beings, a departure from rationality *simpliciter*. Narrowly viewed, it leads to irrational actions, but overall, memory ought to be compartmentalized, given our limitations, in particular, the slowness of algorithmic search. Global rationality requires some local irrationality.

Just how memory ought to be compartmentalized depends on, as well as the agent's beliefs and desires, various parameters of the psychological mechanisms involved, such as search speeds. The above argument does not establish that the actual is the ideal, that typical human compartmentalization is in fact optimal; it only establishes that some compartmentalization is needed for minimal rationality. To that extent, we have justified the ways of God, Nature, or natural selection and man, to Man.

9. CONCLUSION

With the framework we now have, let us reconsider the tacit and highly idealized model of the agent's memory presupposed in philosophy. Quine's conception of the belief system of the epistemic agent was on the dynamic model, from classical physics, of the equilibrium of a physical system; such a passive, homeostatic model seems very far removed from reality. A more adequate representation would be at least an information-processing model. In terms of the latter type of model, Quine might be described as viewing the entire belief system as contemporaneously fully activated or processed in parallel; the contents of the short-term working memory would be the complete long-term memory. A slightly more realistic hypothesis would be that, while short-term memory capacity is limited, all and only the appropriate beliefs at a given moment are always recalled. Such an ideal retrieval efficiency would require at least that exhaustive search of the long-term memory be feasible. This would still be a fundamentally inadequate model, since the success of an actual search cannot be guaranteed; it is as uncertain as running a maze – that is, proceeding through the network structure of the long-term memory. We thus find several layers of idealization in this typical philosophical model.

Several kinds of problems arise when one employs memory models as idealized as Quine's. First, as we saw, one cannot even make sense of an important and prevalent range of human behavior, involving making obvious mistakes. Second, one cannot understand the ultimate rationality of such lapses, as the cost in a tradeoff for some indispensable benefits. Consequently, one will give unsound epistemic advice. For instance, Keith Lehrer's response in *Knowledge* to the Preface Paradox was that we should not accept the statement F, "At least some of my beliefs are false", because that would assure the inconsistency of our belief set (p. 203). But when we reject the model of an ideally efficient memory, we

can recognize that our belief set is highly likely already to be inconsistent, and we can make sense of this fact and its overall rationality. Therefore, since the cost of adding F to our belief set is actually very small, and F is likely to be true (and it would not be a good use of our resources to try to make F false, nor is it likely we could succeed), we ought to accept F.

In these ways, we have shown some of the philosophical significance of the psychology of memory, in particular, its indispensability for an understanding of rationality for a creature in our basic predicament. 'Feasible Inferences' reached a similar conclusion regarding the indispensability of a theory of the agent's reasoning psychology. Thus, more generally, to ignore the question of the "psychological reality" of one's model of how the agent represents and processes information is to exclude the possibility of a philosophically adequate account of rationality, or of notions such as belief, preference, and meaning that depend on it.

NOTES

* I am indebted to Charles Chihara, William Craig, and Barry Stroud for especially valuable comments and suggestions. Some of this material appeared, in a different form, in my Ph.D, dissertation, 'Beliefs and Logical Abilities', University of California, Berkeley, 1977. The paper was read to the Tufts University Philosophy Colloquium, March 1980.

[1] W. V. Quine, 'Epistemology Naturalized', in his *Ontological Relativity and Other Essays* (Columbia Univ. Press, New York, 1969); A. Goldman, 'Epistemics: The Regulative Theory of Cognition', *Journal of Philosophy* **75** (1978), 509–523.

[2] W. V. Quine, *From a Logical Point of View* (Harvard Univ. Press, Cambridge, 1961), p. 42; see also pp. 43 and 44.

[3] W. V. Quine, *Word and Object* (MIT Press, Cambridge, 1960), p. 11.

[4] There is a second way of interpreting Quine's assertion about the revisability relations among beliefs. Such a "normative" interpretation corresponds to that usually adopted for axiomatizations of epistemic logic and decision theory: The entailment between changing one belief and changing a logically related one may not be recognized and acted upon at all by the agent. According to this interpretation, *if* the agent is to be "rational" at least in some narrowly epistemic sense, and in particular, if he is to maintain perfect consistency, he must (or ought), after rejecting the one belief, to reject the other. Since Quine's assertion then does not involve any prediction that the agent will in fact make the appropriate reevaluation, this interpretation seems to be excluded in favor of the descriptive interpretation by the passages cited above. However, many discussions of rationality, including sometimes Quine's (e.g., in the last section of 'Two Dogmas'), do not always distinguish between normative and descriptive rationality theses. To the extent that they do not, strong idealizations of the agent which might be appropriate as part of a normative account then gain unwarranted plausibility as part of a descriptive account. See §7 of my 'Minimal Rationality', *Mind*, **90** (1981), 161–183.

[5] See, for example, A. C. Hilding's letter in *Science*, **187** (1975), 703.

[6] See R. de Sousa, 'Rational Homunculi', in A. Rorty, (ed.), *The Identity of Persons* (Univ. of California Press, Berkeley, 1976), p. 233; and K. Lehrer, *Knowledge* (Oxford Univ. Press, Oxford, 1974), p. 203.

[7] A typical recent text is M. Howe, *Introduction to Human Memory* (Harper & Rowe, New York, 1970); R. Klatzky, *Human Memory: Structures and Processes* (W. H. Freeman, San Francisco, 1975) surveys several current approaches in psychology of memory. A recent alternative to, or elaboration of, the traditional duplex theory is a "levels of processing" theory (cf. F. Craik and R. Lockhart, 'Levels of Processing: A Framework for Memory Research', *Journal of Verbal Learning and Verbal Behavior*, **11** (1972), 671–684). The nature of the difference between the two accounts needs clarification; the levels of processing theory still seems to include the main element of the duplex theory, a distinction between activated and inactive items.

[8] See G. Miller, 'The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information', *Psychological Review*, **63** (1956), 81–97.

[9] As an example, see A. Collins and M. Quillian, 'Retrieval Time from Semantic Memory', *Journal of Verbal Learning and Verbal Behavior*, **8** (1969), 240–247; a recent text is P. Lindsay and D. Norman, *Human Information Processing* (Academic Press, New York, 1977), chs. 8–10.

[10] The *locus classicus* is F. Bartlett, *Remembering: A Study in Experimental and Social Psychology* (Cambridge Univ. Press, Cambridge, 1932); a more recent influential discussion is U. Neisser, *Cognitive Psychology* (Prentice-Hall, New Jersey, 1967), especially chs. 8 and 11. See also J. Bransford and J. Franks, 'The Abstraction of Linguistic Ideas', *Cognitive Psychology*, **2** (1971), 331–350.

[11] See my 'Feasible Inferences', *Philosophy of Science*, **48** (1981), 248–268.

[12] Another source is the a priori assumption that an inference that is obvious for the observer must be obvious for the subject; see 'Feasible Inferences'.

[13] For Quine's view, see *Word and Object*, ch. 2, especially p. 59. On preference transitivity, see for instance, D. Davidson, 'Psychology as Philosophy', in J. Glover, (ed.), *The Philosophy of Mind* (Oxford Univ. Press, Oxford, 1976), pp. 49–50. (Descartes, however, was aware of intransitivity phenomena; see rule VII of his *Rules for the Direction of the Mind*.) On the *modus ponens* inference, see, for example, M. Black, 'The Justification of Logical Axioms', in his *Margins of Precision* (Cornell Univ. Press, Ithaca, 1970), p. 21; using in effect a portion of the common sense duplex model, R. de Sousa rejects similar positions in 'How to Give a Piece of Your Mind; or the Logic of Belief and Assent', *Review of Metaphysics*, **25** (1971), 52–79 (especially pp. 65, 73).

[14] See the articles in Part I of P. Wason and P. Johnson-Laird (eds.), *Thinking and Reasoning* (Penguin, London, 1968).

[15] W. J. McGuire, 'A Syllogistic Analysis of Cognitive Relationships', in C. Hovland and M. Rosenberg (eds.), *Attitude Organization and Change* (Yale Univ. Press, New Haven, 1960).

[16] For example, Howe, pp. 47, 55; and Lindsay and Norman, p. 351.

[17] H. Selby-Bigge (ed.), (Clarendon Press, Oxford, 1965); see especially pp. 20–24.

[18] See §3 of 'Minimal Rationality'.

[19] A 237, B 296; N. Kemp Smith, (tr.), (London: 1929).

[20] On the general design of efficient data structures, see D. Knuth, *The Art of Computer*

Programming, Vol. 3 (Addison - Wesley, Reading, 1973); for a review of search methods in artificial intelligence, see ch. 4, P. Winston, *Artificial Intelligence* (Addison - Wesley, Reading, 1977).

[21] Herbert A. Simon, *Administrative Behavior* (Macmillan, New York, 1947), pp. 80–81.

Received November 10, 1980.

Department of Philosophy
Tufts University
Medford, MA 02155
U.S.A.