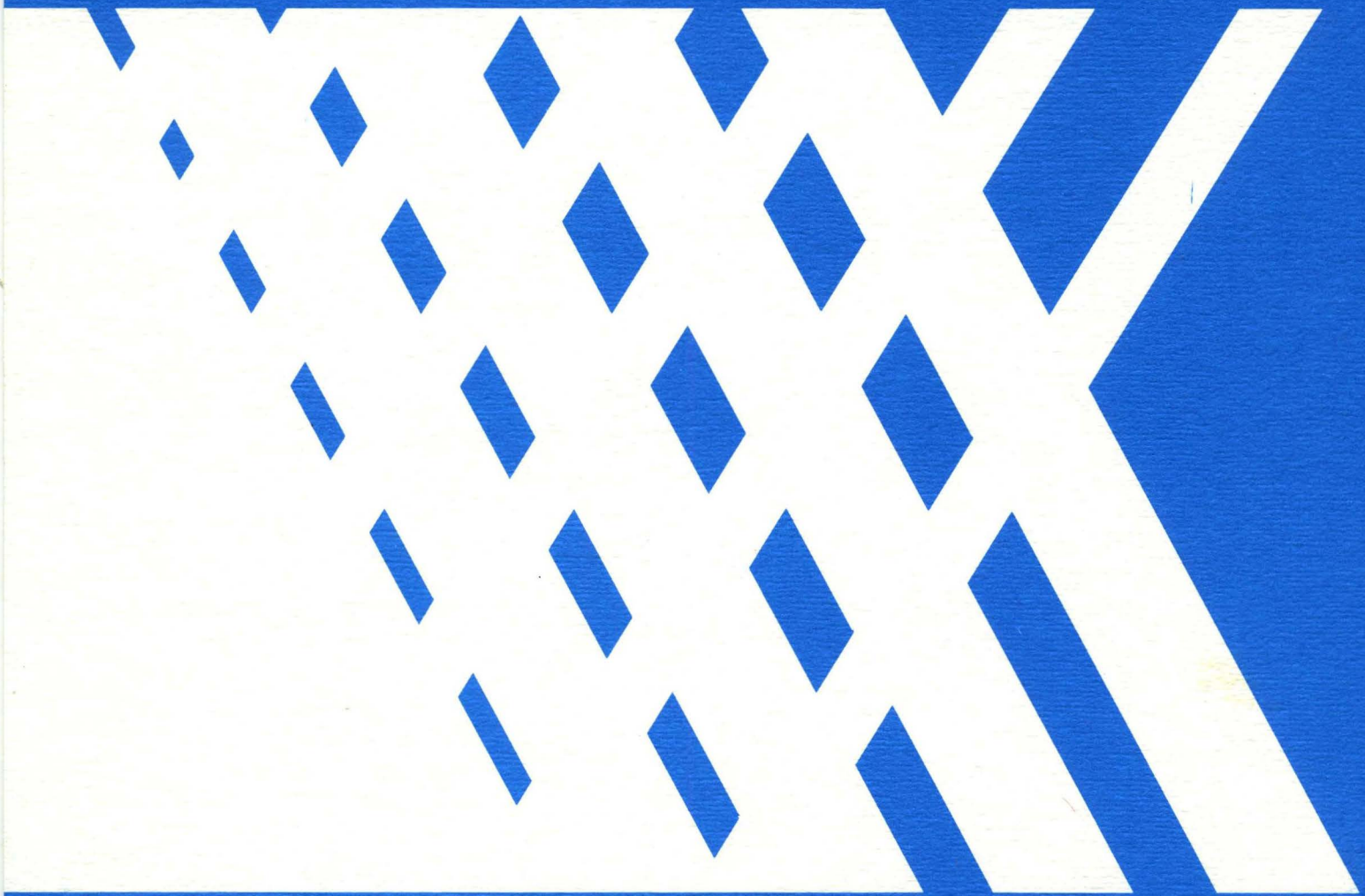
SOME PRINCIPLES OF MEMORY SCHEMATA

BY DANIEL G. BOBROW AND DONALD A. NORMAN



XEROX
PALO ALTO RESEARCH CENTER

SOME PRINCIPLES OF MEMORY SCHEMATA

BY DANIEL G. BOBROW AND DONALD A. NORMAN*

CSL 75-4 JULY 1975

This paper deals with two related issues about memory: access and processing. Consideration of the properties of human memory lead us to suggest that memory is organized into structural units: *schemata.* We suggest that memory schemata refer to one another by means of *context dependent descriptions* that specify the referent unambiguously only with respect to a particular context. We argue that this method of memory reference has a number of desirable features for any intelligent memory system. For one, it leads automatically to metaphorical and analogical match of memory structures. For another, it produces systems that are robust and relatively insensitive to errors.

Consideration of systems which have limits on processing resources leads to some basic principles of processing that apply to memory structures. The quality of output of some processes is limited by the quality of data available to them (these are *data-limited processes*). The quality of the output of other processes is limited by the amount of processing resources available to them (these are *resource-limited processes*). All processes are either data-limited or resource-limited. We suggest that the overall system is driven from two levels--by the data, and by concepts or hypotheses of what is expected. These considerations of processing principles provide some useful interpretations of psychological phenomena, and suggest possible useful computational models for artificial systems.

Key Words and Phrases:

memory, access, schemata, human processing models, context-dependent descriptions, data-limited processes, resource-limited processes

CR Categories:

3.65,3.36

*Donald A. Norman is in the Department of Psychology, University of California, San Diego.

XEROX
PALO ALTO RESEARCH CENTER
3333 COYOTE HILL ROAD / PALO ALTO / CALIFORNIA 94304

TABLE OF CONTENTS

I. INTRODUCTION

A fundamental aspect of the structure of material contained within a large, intelligent memory system is that the contexts in which units of the stored information are accessed are critically important in determining how that information is interpreted and used. There are numerous proposals for the representation of information within memory. Most of the schemes currently under active consideration can be viewed as variants of list structures or semantic network structures. All these proposals have a number of common features, including context-independent linkage between units, and separation of processing and data elements. In this paper we propose a different form for the representation of information which embodies the opposite assumptions about linkage and the separation of data and process. We examine some implications of these memory structures with respect to how the connections among different memory units are formed and interpreted, and we examine some of the issues of processing that arise when these memory structures are used.

The form of our structures is an amalgamation of the principles from the literature on semantic networks, (for example, Norman, Rumelhart, & LNR, 1975; Quillian, 1969) the literature on actors (Hewitt, Bishop, & Steiger, 1974; Kay, 1974) and the new ideas on "frames" (Minsky, 1975; Winograd, 1975). We call our structures *schemata* to emphasize that they differ somewhat from any existing proposals. The word "schema" is taken from the psychological literature, where it has had a long history, most commonly associated with the work on memory by Bartlett (1932), and by Piaget. We propose that one schema refers to another only through use of a description which is dependent on the context of the original reference. We also propose that these schemata are active processing elements which can be activated from

higher level purposes and expectations, or from input data which must be accounted for.

II. MEMORY ACCESS USING DESCRIPTIONS

An important property of human memory is the propensity to find analogical or metaphorical references. One event tends to suggest other events. Sometimes the relationships among the two events are, at best, metaphorical. Sometimes, only some limited aspect of the one event is related to the other. The nature of memory retrieval in humans is, of course, not well understood. We have no hard evidence on the paths followed in the effort to retrieve a particular piece of information or of the sorts of events that one is reminded of while experiencing or remembering another. Despite the lack of firm evidence, we think it important to study memory structures that provide these flexible referential properties. Our goal is to specify a memory structure that allows one schema retrieved from memory to suggest others that should also be retrieved, and that is so constituted that it yields analogical and metaphorical retrieval as a fundamental mode of its operation.

In this paper we speculate on the nature of memory reference processes that can lead automatically, without particular effort, to the richness of the retrievals that we believe to be a fundamental property of human memory. We suggest that memory units refer to one another through the use of descriptions. One memory schema refers to another by describing the other, perhaps by means of a short list of properties of the other. There are different levels of descriptions possible. At the one extreme, a description can be so complete that it unambiguously specifies a unique memory referent. At the other extreme, a description may be so vague that it fits almost every memory referent. We suggest that descriptions are normally formed to be

unambiguous within the context in which they were first used. That is, a description defines a memory schema relative to a context. In novel contexts, a description yields novel results. We call such descriptions *context-dependent descriptions.*

A. Context-Dependent Descriptions

A context-dependent description needs only to be sufficiently precise to specify the desired referent with respect to the context in which it is used. A description contains the important properties of the information relative to some context. This reliance on the power of context is perhaps the most important aspect of retrieval through description. It means, in essence, that a retrieval mechanism must use two sources of information in determining the referent that it seeks: the description and the context. The context delineates some restricted set of elements within the memory that are relevant to the situation: we call these elements either the *focus elements* or the *focus schemata.* The description selects a set of possible candidates from the focus elements. In the ideal case, the description selects a unique candidate element from the focus elements which is the referent being sought. In other cases, there may be no candidates or several candidates, and special processing must occur to resolve the difficulty.

Examples of the combined power of partial descriptions and context are readily available from consideration of perceptual phenomena. For instance, cartoon drawings rely heavily on the fact that although the lines and marks on the paper only provide suggestions (or partial descriptions) of the intended objects, the context created by the overall drawing makes the interpretation of those lines and marks possible. The retrieval mechanism must be designed to cope with close mismatches and with multiple matches. Most likely, it

should operate by attempting to return a single schema in response to a request, even if several schemata were possible, or even if no single schema satisfied all of the description (as long as the violations were not severe). In using an old schema in a new context, this best-match strategy allows identification of analogic similarity.

Consider an example. Suppose that a particular event is witnessed, say that of a large dog (Spot) in a fight with a smaller, relatively weak one (Rover). Given the contextual setting in which the only objects of note are the two dogs Spot and Rover, the description of the scene could be simply a formalization of the statement that two animate objects are present, and the smaller one attacks the larger. Obviously, this is not a complete description; it relies heavily on the contextual setting. In this context, we may be told to associate the term "underdog" with the object in this situation which has the description "small, animate" (that is, Rover). At some later time, if the original setting is retrieved (which identifies Spot and Rover as the only animals in the scene), then this relatively minimal description uniquely identifies the role of each participant.

Suppose now that a new setting occurs, say one with a small person in a fight with a larger one. It causes the perceiver of the scene between the two people to be reminded of the earlier scene between the two animals. The schema for the fight from the earlier scene is directly applicable in this context because of the minimal description used to refer to each participant in the schema. Because this schema is used, the term "underdog" is linked to the smaller person by direct association, in what might otherwise be considered an analogic match. To recognize where the association was derived from, the complete earlier setting would also have to be retrieved, and the perceiver would have to recognize that the old description which applies to

the new setting was derived from this different setting, one with dogs as the characters instead of people.

A single description can apply in many contexts. Thus the minimal description of the fight between the dogs is useful even in inappropriate contexts. Suppose we had the situation of Don Quixote attacking a windmill, or a single individual making a loud, public attack on a large corporation; it is still desirable for the retrieval mechanism to be sufficiently flexible that it can match the windmill or the corporation with Spot, and to match Rover with the protagonist. To do this match, the retrieval system would have to relax the restriction that the attacked object be animate.

B. The Form of a Description

One fundamental issue in determining the form of a context-dependent description is that of deciding on the terms that can be used within a description. Presumably a description contains as its members other descriptions as well as some constant terms. A constant is a description which retrieves unambiguously a single schema independent of context. Initially, in building up memory, some absolute reference points in the memory structure are necessary; built-in constants or primitive terms can serve this purpose. Primitive terms probably consist of grammatical case relations, a basic set of primitive operations, measuring operations, and dimensional terms for spatial and temporal representation. The sensory systems must each contain their basic dimensional primitives, and there must be primitive terms for the concepts and acts that are known by or that can be performed by the system. Once good higher level bases (descriptive terms) for constructing descriptions have been created, descriptions can consist of only non-primitive terms. The major implication of this possibility is that

the "style" of encoding (the choice of terms used) must be reasonably consistent. If not, there will often be the "paraphrase problem" of deciding on the equivalence of two nonidentical descriptions. We believe that such style conflicts are rare within coherent sets of schemata, and that therefore reduction of descriptions to primitive elements is not necessary, nor even usual. Style conformity is aided by a fundamental mode of forming a description. We believe that many descriptions are formed by identifying the schema sought as an instance of another schema, with the new one further specified, or with some changes or exceptions. The *isa* link of semantic networks, and the *beta* notation of Moore & Newell (1973) are examples of the formation of a description (or a schema) by identifying it with another, with certain specified differences.

C. Properties of Context-Dependent Descriptions

Use of partial description and context for reference provides a number of features which we feel are important in a memory system. These are:

* *Efficiency.* Because context is used as part of the address specification, the descriptions within a schema can be short and efficient, providing only enough information to distinguish the referent in context.

* *Generalizability.* A description makes a schema into a generalized form. The same schema can be used in different contexts without changing the descriptions. In the new context, the descriptions contained within the schema will refer to memory structures which have the same relative properties with respect to the new context as the originally intended memory structures had to the original context. Thus memory access by context-dependent description automatically makes a unique, particular schema into a

generalized schema whenever the context is changed. Metaphorical and analogical use of schemata becomes a direct result of the representational scheme and does not require any special mechanism. (In fact, it requires special mechanisms to prevent analogical and metaphorical extension.)

* *Approximation.* A description used in a novel context allows for close matches to be retrieved. Close matches can focus attention on errors or on significant differences in the context from the original.

* *Reliability.* A context-dependent description allows for graceful degradation of function in case of error either in processing structures or in memory structures. Because descriptions are relative, and because the system is designed to cope with descriptions that yield only partial matches or ambiguous matches, any error that produces these results can be handled smoothly. An error in description or process is treated simply as a case of analogical or metaphorical match, and when a failure to match a description occurs, the system can still return with the best possible match (plus a statement of the mismatching aspects).

* *Currency.* Context-dependent description automatically provides a mechanism for referring to the latest version of information. As long as newly acquired information fits a previously determined description, the new information will be retrieved whenever the old description (and the appropriate context) are invoked. This updating requires no change to either the old information or the old description. Of course this implies that a new item with a description similar to an old one will interfere in the retrieval of the old one (and vice versa).

* *Partial knowledge.* Even if the description and context are insufficient to specify the referent, some knowledge is still available. First, it is

apparent in the schema that there is a referent. Second, some aspects of the referent are known and can be used by the system. Finally, a memory procedure can be set in operation to monitor memory for the appropriate referent. As processing continues and as the information relevant to the development of the contextual setting accumulates, previously uninterpretable referents may suddenly become obvious.

III. PROCESSING STRUCTURES

A data structure that is built around schemata which use context-dependent descriptions has a number of implications for the processing structure within which it is embedded. The retrieval mechanism must be reasonably powerful, for it must combine the information from both description and context to determine the set of possible memory schemata relevant to the situation. In addition we propose that each schema is a self-contained memory structure, capable of performing operations because it contains procedural definitions of its potential functions and operations. In general, we conceive of a large number of these active schemata all operating concurrently in a supportive environment, each drawing computational resources from some central pool, each receiving inquiries and generating messages. At this point, we need to specify the form of this operating environment and the general principles of processing that we believe necessary.

Consider the human information processing system. Sensory data arrive through the sense organs to be processed. Low-level computational structures perform the first stages of analysis and then the results are passed to other processing structures. In the awake human, high-level conceptual activity is normally always in progress. Incoming sensory information is either

assimilated into the ongoing cognitive processing or causes an interruption of the ongoing processing. There is a large literature on various aspects of psychological processing relevant to our discussion. Basically, the literature from experimental psychology on attention indicates that the central high-level cognitive mechanisms have a limited processing capacity. When a person is concentrating intently on one demanding activity, that person is able to do very little processing on other activities. Depending on the nature of the tasks, there are rather severe limits on the nature of the activities that may be carried out at the same time. In general, one does not err much by assuming that only a single high-level cognitive task can be performed at any given time. Time sharing of two unrelated tasks, if possible, must be performed with a reasonably slow switching rate, with the time spent on each task between switching measured in seconds.

The limit on the number of activities performed at once does not imply that sensory inputs which are not attended to are ignored, however. Some types of sensory information can be processed, even when the system would appear to be fully loaded. Simple signals can always be detected and complex signals may be detected, although not always recognized. "Important" signals, however, are often capable of attracting attention away from the ongoing activity. The classic examples in the literature are the observations--which have been experimentally confirmed--that although people may be so busy at a task that they claim not to "hear" words or sounds directed at them, they will frequently respond if their name is spoken or if the word fits into the context of the sentence which they are processing at the moment. The problem that these observations pose for theoretical psychology, of course, is that when the importance of a signal is measured by its semantic content, then importance cannot be determined unless the word has been processed.

If, however, all words are processed deeply enough to determine their meaning, what does it mean to be so busy performing a task that "nothing" else is attended to? The psychological literature suggests several mechanisms for handling this problem: in this paper we present a new proposal for this and related problems.

A. Basic Processing Principles

Our analyses of the properties of psychological data and phenomena suggest to us that the human processing system has a number of fundamental principles which underlie its operation, specifically:

* *The processing system can be driven either conceptually or by events.* Conceptually driven processing tends to be top-down, driven by motives and goals, and fitting input to expectations; event driven processing tends to be bottom-up, finding structures in which to embed the input.

* *All the data must be accounted for.* This implies that incoming signals require processing at some level. Thus, a schema to account for a clock's ticking will accept a tick with no further processing demands. If a tick is not heard at the expected time, this is also a datum that must be accounted for.

* *There is a limit to the processing resources available to the organism.* This limit may vary with arousal, but in situations requiring performance on more than one task, each can be allocated only a fraction of the then available resources.

When the resource limit principle is combined with the preceding two, it accounts for interesting aspects of processing behavior. The limited ability to perform several tasks well simultaneously occurs when the resource require-

ments of the data-driven tasks exceed the limit on processing capacity. Ongoing processing is interrupted even when the system is heavily loaded with other tasks because all the data must be accounted for. When processing demands exceed the available limit, a deterioration of performance results. The deterioration usually takes place gracefully, however, and not abruptly. This point is elaborated on in the following section, and is treated in depth by Norman & Bobrow (1975).

B. Data-Limited and Resource-Limited Processes

Here we summarize briefly the points made in more detail in the paper by Norman & Bobrow (1975). When two (or more) processes use the same resources at the same time, they may both interfere with one another, neither may interfere with the other, or one may interfere with a second without any interference from the second process to the first. The important principles are that a process can be limited in its performance either by the amount of available processing resources (such as memory or processing effort) or by the quality of the data available to it. Competition among processes can affect a resource-limited process, but not a data-limited one.

Consider the problem of performing a complex cognitive task. Up to some limit, one expects performance to be related to the amount of resources (such as psychological effort) exerted on the task. If too little of some processing resource is applied, say because processing resources are limited by competition from other tasks being performed at the same time, then one would expect poor performance. As more resources are applied to the task, presumably better and better performance will result. Whenever an increase in the amount of processing resources can result in improved performance, we say that the task (or performance on that task) is *resource-limited.*

Now consider the problem of performing some simple task, say of identifying a sound which is embedded in noise: the processing is limited by the quality of the data. Once all the processing that can be done has been completed, performance is dependent solely on the quality of the data. Increasing the allocation of processing resources can have no further effect on performance. Whenever performance is independent of processing resources, we say that the task is *data-limited.* In general, most tasks will be resource-limited up to the point where all the processing that can be done has been done, and data-limited from there on.

Operations which share the same limited capacity mechanism will not interfere with one another until the total processing resources required by all exceeds some maximum. Moreover, in any given range of resource allocation, one process may interfere with others, but the others need not interfere with it. Just what kind of interference effects are found depends on the particular form of the performance-resource function for each process. Interference can only be observed when a process is operating within its resource-limited region.

Note, therefore, that the effects of interference need not be symmetrical. If task A interferes with task B, but not the reverse, then it would be incorrect to conclude that one of these tasks does not require processing capacity from the same central pool as the other. On the contrary, interference in either direction implies that both tasks draw resource from the same common pool. The asymmetry in effect results when one task is data-limited while the other is resource-limited. Wherever two tasks show an asymmetry in interference effect, it should be possible to demonstrate interfering effects on both by a sufficient change in the availability of processing resources.

One can change the available resources either by increasing or reducing the demands of existing tasks or by adding or removing tasks. Some caution must be used in deciding whether or not one has managed to change resource allocation. If some data-limited task requires some minimum resource *R-min* to operate at all and then operates at its best performance, then the only way to change its demand on resources is either to remove it or to add it anew: no partial allocation of effort is possible.

C. Event Driven Schemata

Schemata are event driven. By this, we mean that all input data automatically invoke processing. These input events must be accounted for. Such inputs generate descriptions which are fed to a number of potential contexts of interpretation, some of which may be suggested by the descriptions themselves. If a quick match is found, the sensory input is fit into a context. The context may itself be a nonprimitive sensory construct whose description might allow it to be fit into a higher level context schema. Associated with a schema may be procedural information which indicates an action to be taken if an instance is found. Such action may demand only low-level responses (for example, having seen a desk, be prepared to see a chair), or may request full use of the central processing facilities ("why was my name said?"). Other internal events can also invoke automatic processing. The recognition of a familiar object in unfamiliar surroundings may trigger special actions.

The amount of processing actually done for a request is, of course, mediated by the total processing load on the system. Schemata that are invoked by sensory events usually cause only low-level decision processes to occur, but the more conceptually based the required decision process, the more

processing effort required: here, the resource limitations can severely limit the performance capability.

Consider an example. In driving an automobile while deeply engaged in some other activity--perhaps talking, listening to a conversation, or thinking--the amount of processing effort left over for the driving is much reduced. Driving, however, is a task that automatically creates a continual flow of new sensory signals, and these sensory signals usually demand low-level processing sufficient for the driving to be done; in general we cannot so distract ourselves by an interesting alternative activity that we entirely neglect the relevant driving activity. Although the mechanics of driving can take care of themselves at a low-level (this is true of most over-learned event-driven activities), higher level cognitive aspects of the driving task are not usually event driven, and they will suffer. Thus if too deeply engrossed in other tasks, the overall level of driving will suffer. No planning activity will take place. An impending decision point may not be anticipated, so that braking and steering activity will take place only when the sensory signals require them, not at an early enough time to ensure smooth, high quality performance.

An important feature of our proposed processing strategies is that, although all the data must be accounted for, it does not really matter how. We believe that there is sufficient flexibility in the use of schemata that an incorrect or very general accounting for data does not cause harm. When sensory events are misinterpreted, for most purposes it will not matter, if only because we simply do not care about most sensory events. For most purposes the original interpretation is quite adequate. When better interpretations are needed, then the schemata can be expanded or modified to provide them. Initially, all the data must fit into some schema, but it does not matter if the fit is bad.

D. Depth of Processing

Everything that arrives at the organism must be processed to some extent. Because the processing resources of all devices, including the human, are finite, there must be something that distributes the processing resources that can be allocated to any task: there must be some scheduling device.

What things should be processed in depth? We argue that it is most important to process what is least expected. If an event occurs that is totally expected, then there is little information to be gained from its detailed analysis. If the event deviates from expectations, or if an event that is expected fails to occur, or if an event that one is not prepared for does occur, then these are special events and must be given priority in processing. Thus it is that the things that we most expect to see or experience will leave the least impact on us: it is the discrepancies that we will note. Moreover, the same basic principle tells us how much to process discrepant events: we process them until we know how to account for them. At that point, they are no longer discrepant and, therefore, no longer need processing.

When we say that "all the data must be accounted for", we mean simply that some conceptual schema must be found for which these data are appropriate. If the data are seen not to be of importance to the central analyses of the moment, then almost any schema will do. If the data appear to be important--and this importance is determined by the nature of the schema for which they appear to be relevant--then processing in depth will probably be necessary to elaborate on the manner in which those data are interpreted beyond that provided by the initial schema. Finally, if the data cannot readily be accounted for, then we suspect they create an interruption in the processing cycle, for they will demand sufficient resources from the system to enable them to be processed sufficiently to be understood at whatever level is necessary.

The psychological literature on memory indicates that events that are not processed deeply are not well remembered: the deeper processing, the better the memory for those events (see Craik & Lockhart, 1972). One would expect, therefore, that data which were readily accounted for would not require much processing, would not be well remembered, and probably would not receive any conscious attention. Data which either were deemed to be important or which could not easily be accounted for would, however, receive sufficient processing effort and, as a result, they would probably be remembered later. Moreover, we suspect that they would receive conscious attention at the time of their arrival and processing. Thus data which are expected or otherwise readily accounted for would be ill remembered. This would help explain why we need not be concerned with every detail or anomaly of the environment. To use an example provided by Abelson (personal communication), a red stain (tomato) on the manuscript copy of this paper could be accounted for by low-level organizational schemata (namely, the schema for stains and shadows). We would not necessarily even be aware of the stain, despite the fact that a reasonable amount of processing effort was expended in accounting for it. It is only if the stain could not readily be accounted for that it would reach conscious awareness (as would be the case, say, if the stain would move about on the page or float one inch above it). Events which are very close to expected events may also be assimilated to their expectations. In this case, the differences will probably not be remembered, only that the general schema was instantiated.

E. The Organization of Processing

We view the cognitive processing structure as one that consists of a multilayered assemblage of experts. Each expert is a process that knows how

to handle the data and suggestions provided it. When situations arise that an expert cannot handle, or when communication with the other experts that it knows about fail, then it passes on its information and messages to higher level processes. The entire system consists of a multiplicity of hierarchies of experts, each expert working on its own aspect of processing, interpreting and predicting the data which are available to it, shipping requests to higher processes, and expectations of inputs to lower ones.

An important aspect of the organizational structure of processing concerns the interactions of conceptually driven and data-driven schemata. If the system is to have any function at all, there must exist several overriding considerations. There must be purposes to activities. There must be some procedure for selecting from among all the various activities taking place at any moment those that are most important for the purpose of the system. Basically, we believe that the system must be provided with motivations to provide top-down drives, a capacity to learn, and the ability to be aware of itself. We conclude that there are reasons to postulate a single central mechanism having many of the properties ascribed to human consciousness.

Purpose. Purposes add direction to the system--the top-down hypothesis driven aspects. The principle that all the data must be accounted for adds the bottom-up drive. Both would seem to be essential. Without purpose, the system will fail to pursue a line of inquiry in any directed fashion. Purposes should be at a high level, not local, simple goals. A high-level purpose coupled with sufficient operating principles should thereby automatically produce the necessary subgoals for the immediate demands of processing, and provide criteria for allocation of resources to event driven schema relevant to the purpose.

Motivation. A person can pursue several purposes at one time. One can, for example, be driving home, trying to find a good music station on the radio, and having a conversation. When the demands from these three tasks exceed the capacity of the processor, criteria from the individual purposes cannot mediate conflicts for demands that have differing purposes. There must be a central motivational process which serves this function for most conflicts.

Retrieval and evaluation. The distribution of computing resources should be guided by the principle that all the data must be accounted for: effort is spent processing data that do not fit into any active schema. The existence of data that do not fit an existing schema, or the absence of important data that are required by a given schema are both capable of requesting some central mechanism to examine the nature of the mismatch. The retrieval mechanism must be capable of the evaluative role that must be performed in assessing context-dependent descriptions. Descriptions must be combined with context, allowing metaphorical or analogical retrieval to take place and to be used to useful purpose. We believe that some central mechanism that has access to many memory schemata is essential in performing intelligent evaluation whenever a memory schema has proposed an unsatisfactory match for a description.

A central mechanism. We believe that all these considerations together require that the system be guided from the top by a single central mechanism, one with awareness of its own processes and of the information sent to it by lower order schemata. We believe this central conscious mechanism controls the process that schedules resources, initiates actions by making decisions among the alternatives presented to it, and selects which

conceptualizations to pursue and which to reject. We assume that this mechanism keeps track of its operations and of the overall context by means of a small capacity memory structure, probably the short-term memory structures that are widely discussed in the psychological literature. We believe this central evaluating mechanism is probably serial, probably slow, and probably resource-limited. One major argument for the existence of a single, central control mechanism is that despite the multiplicity of processing structures, there is only one body. There must be coherence and unity in the overall control. Conflicts must be resolved and important decisions must only be made once. These statements do not mean that there cannot be several high-level mechanisms, each specialized for certain types of decision or control functions, each perhaps having different modes of operation. The important point is that at any specific time, for any given task or for any given process, only one of those mechanisms must be in control at any moment.

IV. SUMMARY

In conclusion, we propose that memory structures be comprised of a set of active schemata, each capable of evaluating information passed to it and capable of passing information and requests to other schemata. We suggest that a memory schema refers to others by means of a description that is only precise enough to disambiguate the reference within the context in which the original situation occurred. This context-dependent description thereby provides an automatic process for creating general memory references from specific events, allowing for automatic generation of analogical or metaphorical memory matches. The retrieval mechanism that operates upon the descriptions must be intelligent enough to combine both descriptions and

context in a meaningful, useful manner, and it must be relatively insensitive to mismatches and underspecification.

The processing structure of the memory system is one that has a limit on resources that are available Any given process is either data- or resource-limited. Some scheduling device is necessary to keep things operating smoothly. We believe the system to be driven both by the data (in a bottom-up fashion) and conceptually (in a top-down fashion). The principle that "all the data must be accounted for" guides the bottom-up processing. We believe that a single, conscious high-level mechanism guides the conceptual processing, taking into consideration the motivation and purposes of the organism.

Conscious processes are invoked whenever underlying schemata provide information for evaluation, whenever new processes must be invoked or old ones terminated, or whenever the output of one schema must be communicated to others not immediately invoked. Any time that there is a mismatch between data and process or expectations and occurrences, conscious processes are brought in. The automatic, active schemata of memory and perception provide a bottom-up, data driven set of parallel, sub-conscious processes. Conscious processes are guided by high-level hypotheses and plans. Thus consciousness drives the processing system from the top down, in a slow, serial fashion. Both the automatic and the conscious processes must go on together; each requires the other.

ACKNOWLEDGEMENTS

This paper was written while Norman was a fellow at the Center for Advanced Studies in the Behavioral Sciences, Stanford, California, and we are grateful to the Center for the facilities which they provided. Research support to Norman was provided by grant NS 07454 from the National Institutes of Health. It will appear as a chapter in a book of studies in cognitive science (Bobrow & Collins, 1975).

REFERENCES

Bartlett, F. C. *Remembering: a study in experimental and special psychology.* Cambridge: Cambridge University Press, 1932.

Bobrow, D. G., & Collins, A. *Representation and Understanding: Studies in cognitive science.* San Francisco: Academic Press, 1975.

Craik, F. I. M., & Lockhart, R. S. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 1972, *11*, 671-684.

Hewitt, C., Bishop, P., & Steiger, R. A universal modular ACTOR formalism for artificial intelligence. *Proceedings of the Third International Joint Conference on Artificial Intelligence*, 1973, 235-245.

Kay, A. *SMALLTALK, A communication medium for children of all ages.* Palo Alto, California: Xerox Palo Alto Research Center, Learning Research Group, 1974.

Minsky, M. A framework for representing knowledge. In Winston, P. (Ed.), *The psychology of computer vision.* New York: McGraw-Hill, 1975.

Moore, J, & Newell, A. How can MERLIN understand?. In Gregg (Ed.), *Knowledge and cognition.* Baltimore, Md.: Lawrence Erlbaum Associates, 1973.

Norman, D. A., & Bobrow, D. G. On data-limited and resource-limited processes. *Cognitive Psychology*, 1975, *7*, 44-64.

Norman, D. A., Rumelhart, D. E., & the LNR Research Group. *Explorations in cognition.* San Francisco: Freeman, 1975.

Quillian, M.R. The teachable language comprehender. *Communications of the Association for Computing Machinery*, 1969, *12*, 459-475.

Winograd, T. Frame representations and the declarative-procedural controversy. In D. Bobrow & A. Collins (Eds.) *Representation and Understanding.* San Francisco: Academic Press, 1975.