ONR Final Report 1988

Marvin Minsky
August 1988

This project was concerned with developing a theory of intelligent thinking and learning, based on the "Society of Mind" model of intelligence. This research was funded over a period of years by the Computer Science Division of the Office of Naval Research. It involved not merely the past three years but the entire period from the early 70s when the Society of Mind hypothesis began to take form. In 1987 we published a major book on this theory: *The Society of Mind*, Simon&Schuster, followed by a paperback version in 1988 and translations into many languages. Our research included the following specific subjects:

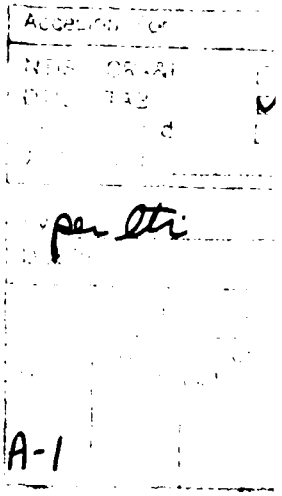
- Connectedness of Parallel Computers
- Exploiting Parallel Processing
- Connectedness of Commonsense Knowledge Bases
- Connectionism and Society of Mind
- Advantages and Deficiencies of Connectionist Networks
- Insulation and Interaction
- Learning and Representation
- Intermediate Units and Significance
- Associations and Connections
- Unifying Frames and K-lines
- Clarifying Conceptual Dependency
- Computational linguistics
- Research tools for society of mind models
- Discovery processes
- Bridges between symbolic and connectionist models

Connectedness of Parallel Computers

It has recently become feasible to build computers with many thousands of small processors, and soon it will be practical to have several millions. Will further increases in the size of such machines limit their usefulness by requiring increasing numbers of connections per processor? We believe that this will not be a serious problem in the areas of systems that use very large bodies of commonsense knowledge because, according to our theory, such systems will naturally tend to be partitioned into domains or "agencies" with relatively weak demands for intercommunication.Accordingly, it should be feasible to assemble large such machines from smaller ones by using relatively few additional wires. Indeed, we conjecture that in typical applications the required numbers of connections per processor will approach a fixed and practical bound.

In The Society of Mind we conjecture that commonsense reasoning systems need not increase in connectedness as they grow in size and complexity, because they will tend to evolve into clumps of specialized agencies, rather than homogeneous networks. Indeed, because those sub-systems work best with different, specialized types of internal representations, it is neither necessary nor practical for one of them to communicate directly with the interior of another. Furthermore, because most acquired skills will evolve from older ones by differentiation and specialization, this will bias the largest scale connections to evolve into the form of tree-like (as opposed to network-like) arrangements.

If a mind is assembled of distinct agencies with so little inter-communication, how can those parts cooperate? What keeps them working on related aspects of the same problem? The answer proposed in The Society of Mind is that cooperative activity is less important than "exploitative" activity. Because these specialized agencies use different internal languages and representations, they cannot understand one another, and this means that each of them must learn to exploit some of the others for their effects - without knowing how those effects are produced. To be sure, this requires yet other agencies to manage those assortments of specialists; otherwise the system will be subject to conflicts about access to limited resources. Furthermore, those management agencies themselves cannot directly deal with all the small interior details of what happens inside their subordinates. They must work, instead, only with summaries prepared by these subordinates. Without such constraints, the managers themselves will be overwhelmed. And this argument generally applies recursively to the insides of all those agencies, hence relatively few direct connections are needed except between adjacent "level bands".



Exploiting Parallel Processing

The Society of Mind theory is ideally adapted to parallel processing because it is based on large numbers of simple agents that all use similar processes. Indeed,

several students are adapting it to the "Connection Machine" type of parallel computer based on an architecture in which large numbers of processors execute nearly identical programs. Although that can constitute a serious restriction on parallel programming in general, it is much less limiting for systems in which most of the knowledge is represented less in the processes themselves than in their network of interconnections. In this sense, our theory is related to - and builds bridges between the older "semantic network" models in AI and the newer "connectionist," "neural network," and "data level reasoning" models now being being explored worldwide.

Present day hardware technology makes SIMD machines particularly feasible. However, despite this potentiality, many computer scientists have been repelled by their apparent limitations. We expect SIMD systems based on our theory to be able efficiently to support the computations required to build machines able to apply commonsense reasoning to large, heterogeneous knowledge bases. (It is no accident that Society of Mind mechanisms can be supported by the CM type of machine; both that theory and that architecture developed in the same environment.)

Connectedness of Commonsense Knowledge Bases

We conjecture that the assemblies of representations which interconnect the clumps of expertise inside each person's mind must form a weakly connected network. Unfortunately, we have no formal theory, yet, of how and when large commonsense reasoning systems will need relatively fewer connections? As I see it, the problem is that most present-day theories of computational complexity are based either on worst-case analyses or on statistical arguments - and neither approach well represents practical reality. The worst-case theories emphasize only potentially intractable aspects of problems which, in their usual forms, present no practical difficulties; the statistical theories tend to assign equal weights to all instances, for lack of systematic ways to emphasize situations likely to be of practical interest. The AI systems of the future, like their human counterparts, must be evolved to "satisfice" rather than optimize - and traditional complexity theories were not designed for such requirements.

This phenomenon reminds me of Shannon's proof that, in general, Boolean functions of several variables require switching circuits that grow exponentially with the dimension. Again, in practical experience, that never seems to be the case. Why not? Because functions that are "interesting"- that is, functions which "make sense" in one way or another - seem always to be composed in simple ways, from simpler functions, presumably because we can make complex new ideas only by combining simpler ones. Functions that lack this character tend to be incomprehensible, and we can conceive of few uses for them. Shannon's proof shows that, as n increases, the proportion of functions that can be described as compositions of simpler ones grows exponentially small. These

tend to be the "meaningful" functions - but are just the ones that slip through the net of such theories. To the extent that this is a sound analogy, the germ of a technical foundation for theories about the structure of "meaningful" knowledge might already exist in the algorithmic complexity theories of Solomonoff, Kolmogoroff, Chaitin, Levin, et al - because those theories do indeed assign more weight to the most composite functions; indeed, this has been proposed as a basis for learning, induction, and common sense inference. Unfortunately, no one has yet found practical application of those theories, but the recent work of Rivest and his students suggests that there may be a way.

Connectionism and Society of Mind

Why is there so much excitement about connectionist research today? Some researchers simply want machines to do the various sorts of things that are usually called intelligent. Others hope to understand what makes people able to do such things. Yet other researchers yearn for ways to avoid writing computer programs: how much more pleasant it would be if we could build, once and for all, machines that learn to improve themselves. Then, whenever we want to have something new, we would simply explain - or demonstrate - what we want, and let those machines attempt their own experiments, or read some books, or go to schools ; that is, to do the sorts of things that people do. Why can't we make machines like us, that grow by learning from experience?

How can we make machines that can learn? One approach is to start at the top, at the level of commonsense psychology: try to imagine the processes by which a person plays a certain game, solves a particular kind of puzzle, or recognizes a specific sort of object. Then if you cannot find a single, simple way to do such a thing, try to break down those processes into simpler parts that you can connect together - either in hardware or in software. This so-called "top down" strategy is typical of the approach to AI called heuristic programming, which has developed productively for several decades.

To go in the other direction, we can use a complementary strategy. Begin with parts we already understand, and work upwards in complexity to find ways to interconnect those smaller parts to accomplish our larger scale goals. We can start with almost anything - with small computer programs, elementary logical principles, or simplified models of what brain cells do. This "bottom-up" type of strategy is typical of the approach to AI called connectionism. It has only recently started to flourish, despite a long history of successful use on smaller scales, in various forms of adaptive network systems.

Why did the field of connectionist research start to expand only so recently - considering that connectionistic models have been the dominant psychological theories for more than a century. One answer is that in the early days of computers, heuristic methods developed so quickly that connectionist networks were swiftly outclassed. Connectionist experiments required prodigious amounts

of computation that only became available over the past few years. But there were other, more fundamental reasons. We discuss this in detail in the first and last chapters of the 1988 edition of the book *Perceptrons*.

Which approach is best to pursue? The answer is simple: we must use both. In favor of the top-down side, research in AI has told us much about how to make machines solve problems by using methods that resemble reasoning. In favor of the bottom-up approach, the brain sciences have told us a little (but only a little) about what brain cells do. If we knew more about brain cells and their connections we could use that knowledge to work directly toward discovering how they support our higher level processes. If we understood more about human psychology we could work toward finding out how brain cells do it. But right now we're caught in the middle; we know too little at either extreme. The only practical present option is to ping-pong between them, making theories for building plausible bridges. How can we do that? One way is to focus on inventing various ways to represent knowledge, and then to try to extend those techniques in both directions. On the connectionist side we can try to design neural networks which can use - and learn to use - various types of *representations*. On the *top-down side*, we can try to design higher level systems that can effectively exploit the knowledge thus represented. This is basically what we have tried to do in our Society of Mind research project.

Advantages and Deficiencies of Connectionist Networks

The dream of the early connectionists was to start with virtually nothing at all but a loosely connected network of parts that would somehow be able to learn by itself; it was hoped by some that, once some modest goals were so achieved, little more might remain to be done than enlarge those miniature networks to have enough capacity to learn to become intelligent! Several such systems were actually built and they learned to accomplish various things - but none of those systems got very far. Why might one have expected them to do so much more by themselves? There were, and still are, many reasons why weighted-connection learning machines seem promising. They certainly resemble aspects of what we think we see in brains. We know that they can be designed to recognize many types of patterns. It is clear that their redundancy can make them resistant to noise and injury. And we know some surprisingly simple algorithms that indeed permit them to learn - while automatically discovering certain types of generalizations. In some of the more recent work, neural networks have been shown able to discover certain kinds of knowledge representations spontaneously, without these having been specified in advance. And because the units of those networks all operate simultaneously, they promise to offer the power and speed of genuine parallel computation. In particular, the networks and learning algorithms most

thoroughly studied today appear to be very good at learning to represent predicates that can be defined in terms of the addit ve influence of large numbers of relatively independent inputs. These amount, effect, to patterns that can be recognized by processes that match input vectors to prototype vectors, using continuous, arithmetic matching criteria. This is very important because these are just the sorts of predicates that are very poorly recognized by algorithms of essentially logical character. In short, weighted-connection representations have many virtues, including these:

- The learning algorithms, when they work, can be extremely simple.
- The networks can often automatically find useful generalizations.
- They automatically find good matches in the presence of noise.
- They tend to be redundant and insensitive to injury.
- They can operate with great speed.

Given so many advantages, one might ask why ever use discrete, symbolic schemes at all? The trouble is that all simple connectionist schemes appear to have serious problems of their own. Not very much is yet known about this, but we can conjecture that these problems will become increasingly serious as the experiments are scaled up in size. In particular, we expect to find that when we attempt to make a single, highly connected network learn to accomplish several tasks of different character - that is, problems that require representations of conflicting types, the internal interferences will become worse as the networks are increased in size. Furthermore, it will be very hard to train such networks to accomplish tasks that are basically serial or recursive in character -such as counting the number of, or distinguishing between the features of, the different objects in a picture. This is because such processes require the making of clear separations in memory between things which are very similar. To put it epigrammatically, connectionist networks are good at making connections between things, but are not so good at keeping them separate. Enthusiasts of neural networks will disagree with this, and only the future will tell which view is more realistic. In any case, the use of weighted numerical representations entails a heavy price. Although such methods can yield useful types of performances, they will also be prone to reaching intellectual dead ends. This is because it is in the nature of numerical representations to combine components in an opaque fashion. When you add several numbers, you can no longer recognize, in the sum, the influences of the components that were combined. Consequently, such systems do not lend themselves to be useful as parts of larger, more reflective systems, except as "black boxes" to be used only for their effects. This is because the nonavailablitity of explanations makes it difficult for other systems to explain the results; and then separate out components for constructing useful new variants. For further learning, then, the credit assignment problems

may become intractable. In short, such systems will tend to become unable to do reasoning. Thus, clever pseudo-holistic representations become obstacles to further intellectual growth.

I don't mean to suggest that we always can, or always should, avoid using schemes like that of assigning weights to evidence. The conditions of real life constantly compel us to make decisions that amount, in effect, to ceasing to think. Indeed, since virtually every overt action we make reflects some such decision, the results of weighted comparisons probably have a disproportionate influence on the most immediate causes of our observable behavior.

Insulation and Interaction

We tend to think about thinking in positive terms: of assembling parts into larger wholes by making connections among ideas. But negative connections - call them insulations - are just as important as positive interactions. One might say that insulations build up barriers, while interactions break them down. Too many interactions lead to confusion and inefficiency; too much insulation leads to incoherency. To illustrate the problem in an evolutionary context, consider this example, from *The Society of Mind.*

Imagine a certain animal in which some new mutated gene produces a substance S that comes to play two different, vital roles at once - one in the heart and one in the brain. Now that animal's descendants can do more with less, so natural selection will tend to further improve that gene and disperse it among the flock. But consider the cost of that short-term gain: that double-purpose protein is an obstacle to further improvements in either heart and brain! Any change in S that strengthens one would almost surely weaken the other, because the earlier form of the gene has already evolved to constitute the best available compromise. Now our new, mutated animal is doomed to be very slow to evolve because each further change in S disrupts so many processes. By breaking down the separateness of the mechanisms of the heart and the brain, S constitutes a new constraint that keeps them from learning independently. Each such constraint makes it harder to change - and the short term gain from finding two different uses for S is a long term evolutionary liability. The peak upon which our species is stuck is actually an artifact - produced by causes that interact only through an evolutionary accident that constrains a sum to a smaller result than we could get by separately climbing two different peaks.

How does our evolution ever manage to escape such double-purpose deadlock states? Our very early ancestors evolved a trick that has become an essential part of all of our subsequent evolution. The secret lies in the simple fact that the processes used to copy genes are prone to make duplicate strings of genes. Then, whenever a duplicate gene mutates, its unchanged twin can still perform its original function, while the variant gene can drift along a separate evolutionary track. Two versions of the gene for S could thus then manufacture different

chemicals, one of which can improve the heart, while the other enhances the brain. We usually think in positive terms about making wholes by combining parts. But when interactions lead to inefficiency, we may need negative connections - call them insulations - to keep things from getting confused.

It is not only in connectionist nets that multi-purpose deadlocks can be deleterious. It can happen as well in symbolic realms. Suppose that a certain concept C was learned in a certain context X - but now we want to apply that skill to another, different context Y. If C attempts to continue to learn in both contexts,then, the more that X depends on C's original details, the more X may be handicapped when those details are modified to suit the demands of Y. Again we could try to avoid this by providing Y with its own duplicate of C - but if too many copies learn different things, our system will tend to split into incoherent accumulations of non-shareable information. Conjecture: when a neural network learning system gets stuck, it should help to duplicate the "hidden units" with substantial weights and change one of the copies enough to set it on a new course. Later, remove copies that have not further distinguished themselves. We could use even more elaborate interventions, when variation of connection-weights leads to poor performance of a unit; the learning process might replace that unit by rearranging its negative and positive inputs into a pair of separate recognizers that are connected to a third, administrative unit whose performance can later be independently optimized.

It would be prohibitively expensive to frequently duplicate entire networks. Our decades of experience with serial computers have taught us that there is another way: - to avoid massive duplication by using inheritance-based virtual copies. This is fairly easy to do in systems equipped with enough temporary memory. But we still know very little about how to approximate the virtues of virtual copies in connectionist networks. One way is to use the recording units we call K-lines to remember various features of a network's activation state.

How to combine the advantages of symbolic and connectionist methods? Researchers in the symbolic area have developed good ways to constrain fruitless search. In The Society of Mind, we propose to build systems wherein some networks supervise the experiments that occur inside some other nets. Symbolic AI researchers have developed many powerful ways to represent knowledge - but have not developed good ways to make machines develop good new representations; here connectionist methods will eventually help. Nor are present day symbolic systems good enough at discovering, without external help, particular knowledge to fill their rules and frames and scripts and semantic nets. Connectionist methods could help in all these areas - but not until they get better at search. But first we shall have to develop multi-section learning schemes in which some sections supervise how other sections work and learn. Eventually, as we learn more about such matters, we shall begin to attack the longer range goal of discovering how to enable such systems to construct productive new reformulations of problems.

Learning and Representation

In order for a machine to learn, it must have potential ways to represent what it may learn. What representation should we choose? Over the years researchers in AI have made many theories about this, and done many experiments on knowledge representation schemes, such as those called Semantic Networks, Conceptual Dependency, Frames, Predicate Calculus, Rule-Based Productions, Procedural Representations, and quite a few others. This is not the place to review the features of such representations. Instead we'll contrast all of them with the methods used in connectionist nets. The basic goal in that enterprise is to embody knowledge into the conductivities or weights assigned to connections among a network of nodes. The most common form of such a node consists of a linear part that "adds up evidence" and a nonlinear part that "makes a decision".

In principle we can construct such a network to represent any computable function. Consequently, any of those other types of knowledge representation could be encoded into such a network. In practice, however, the linear, additive aspect of typical connectionist nodes can lead to problems because addition itself is so fundamentally opaque in the sense that once several numbers are added up, one cannot recover, from their sum, the inputs that were thus combined. There is a spectrum of possible ways to deal with this basic problem of opacity; in fact, this can be seen as breaking up the entire field of research on connectionist nets. The problem of opacity becomes increasingly severe as the density of connections grows. When each node connects only to a relatively few others, then we have structures that resemble what AI researchers call Semantic Networks; the elements of those types of knowledge representations are comparatively localized. Some connectionist models are of this highly localized type, while others use networks in which a typical node sums a relatively large number of contributions from different sources; these are called highly distributed. It is important to recognize that we probably need quite different learning procedures for dealing with localized and distributed representations.

In any case, once we can represent knowledge in terms of connection weights, it becomes easy to formulate the problem of learning in terms of various established techniques for "hill-climbing" or gradient ascent. To do this, we merely need to express our evaluation of a network's performance in terms of a single numerical success function. Then the problem of learning can be reformulated in terms of searching to find the maximum value of that evaluation function. The terrain to be searched is simply the vector space of the connection coefficients inside our network. Of course, this problem is simple only in principle, because any strategy based on gradient ascent can fail by getting stuck upon a local, isolated peak whose altitude is relatively insignificant. There simply is no local way to ensure that any such procedure will always reach a global maximum, instead of becoming trapped upon some local feature of topography such as a terrace, ridge, or peak,

Sometimes one can escape from traps by making occasional random jumps - and many people even hold that in this lies the key to creativity. One such strategy is the method called "annealing", which works effectively on certain types of problems. But we still have little insight into which classes of problems can be treated that way; indeed, it is easy to construct examples in which annealing leads to worse results than would come from complete, exhaustive search. Such problems not only arise in AI: they lie at the heart of evolution itself. For example, it seems almost a truism that most mutations are deleterious, but it is important to see why this is so. Whenever we see a live animal, we're seeing a system that is highly evolved: in other words, it is virtually certain already to stand on a local peak! Because much of the nearby territory has already been explored, the present location where it stands is likely to be quite close to the best that lies in the structural neighborhood. Therefore, mutations will tend to be bad because they will naturally tend to undo the work of selection that was already done in the animal's evolution. To be sure, there is always a chance to find a better place by making very large random jumps. But unless we do this selectively, the results can be worse than exhaustive search. Annealing may seem efficacious at first, when applied to systems in random states - but it won't fare so well when applied to systems that are already in more highly evolved states. For the more we've invested in finding this peak, the more will be wasted of what we have learned – as soon as we jump away from it.

Contrary to many recent pronouncements, such methods can only ameliorate, but never can eliminate the types of difficulties that gave rise in the 1950s to the field of heuristic programming. No matter how hard we continue to try to extend the powers of methods based on local search, such methods can take us just so far; that search itself will end up trapped upon some abstract, unknown peak in the strategy space of search machines. Finding a peak is a means, not an end; it rarely is our real goal. Instead of just seeking escapes from traps, we might better use those peaks as clues at a deeper level of analysis. When the problem we're solving is easy enough, it may suffice just to climb its hill. But when our problems are deeper than that, then, in place of simply climbing those hills, a better goal would be, instead, to ask ourselves what causes those hills. Which of those peaks reflect inherent structural aspects of the problems we're trying to solve, and which of them are artifacts of the representations we happen to choose?

Intermediate Units and Significance

The problem of the double-purpose deadlock casts a shadow across the entire realm of representations that are distributed, holographic, or generally non-local or holistic. A Symbolist might oppose a Connectionist with an argument like this.

"Practical representations must employ relatively compact, localized, inter-

mediate units, each of which has some degree of individual significance."

Consider, for example, the intermediate signals or agents called microfeatures in Waltz and Pollack, Massively Parallel Parsing," *Cognitive Science*, 9, 1, 1985. Can we expect these each to carry its own, recognizable significance? The answer is that significance itself is a relation between a thing and an observer. Even inside a human brain, only certain signals will be significant to certain agencies. For example, few signals from other parts of the brain would be lacking in any socially communicable significance. Lower level microfeatures may be significant to correspondingly low level agencies, but this will not be expressible, or even comprehensible, to the higher level agencies that exploit what those systems do. This must apply in general to the hidden units that appear in the literature about parallel distributed processing. See also Jerome Feldman's *Neural Representation of Conceptual Knowledge*, TR198, Univ. of Rochester, C.S. Dept, June 1986.

Similar issues arise in connection with discrete, symbolic mechanisms. Just as we can sometimes solve problems by using massively additive representations whose units have little individual significance, we can sometimes solve other problems by using discrete representations composed of similarly insignificant components; for example, when we represent a composite Boolean function as a canonical disjunctive form. In doing that, one disperses information about the function's internal structure and composition. At both extremes - in representations that are either too distributed or too discrete - we lose the structural knowledge embodied in the form of intermediate-level units. That loss may not be evident, so long as our problems are easy to solve, but those intermediate concepts may be quite indispensable for solving more advanced problems of related kinds. This is because the comprehension of a complex situation often hinges on finding a good analogy, or on composing meaningful variations on a familiar theme. But doing this is virtually impossible, with representations that are too fragmentary, which is precisely what happens both when we use canonical logical forms or when we use linear holographic transformations. Those representations are simply too homogeneous. They have no way to represent the significant parts and relationships. The idea of a thing with no parts provides nothing that we can use as pieces of explanation.

None of this should be taken to mean that it is bad or wrong to build distributed systems. Valuable forms of robustness can emerge from redundant representations, and the use of parallel and shared elements can lead to imaginative reformulations. My point is that unless a distributed system has some potential ability to crystallize out important new sub-concepts and substructures, its ability to learn will eventually slow down and it will be unable to solve problems beyond a certain degree of complexity.

But how could a machine discover useful new intermediate elements? Our idea is to invert the usual view that local peaks are nuisances. Instead, we'll regard them as potential indicators of when and where we need to introduce novel elements. We could even argue that the alternative - of further pursuit

of other, better local peaks may be counterproductive in the long run This is because if we try to find solutions by making compromises across many different contexts, then the weights that are finally assigned to connections will tend to embody average rather than peak performances. What makes this problem serious is that it can lead to accumulations of commitments that may become costly or impossible to repair: then the system will achieve a certain level of performance but then be unable to further develop. In The Society of Mind we cast this in terms of a principle: when agencies of equal rank conflict, don't try to satisfy them both. It's better to abandon both and try to find another one - perhaps by appealing to agencies of higher rank.

Associations and Connections

Any machine that learns effectively must discover connections between actions and outcomes, inputs and outputs, or causes and effects. Before a machine could discover such connections, however, it must have some way, at least potentially, to represent them, at least implicitly. After all, we can't connect abstractions themselves, but only their representations. To be sure, any set of associations could be represented, in some purely mathematical sense, by cataloging high-order correlations or high-order polynomial coefficients. Indeed, Norbert Weiner and Denis Gabor proposed in the 1950s to build learning machines on this basis. However, those schemes turned out to be impractical because of exponential growth.

Any theory of the mind must explain how the brain provides enough connections to make the mind capable of such a wide range of associations. It would require too many wires to connect every agent to every other agent. There is no reason to suppose that any clever coding scheme or holographic principle can get around this; the problem gets increasingly worse as the brain increases in size. Nor has evolution itself found any efficient solution to this problem. The human brain has so many connecting wires that the actual nerve-cells constitute only a fraction of its mass. In The Society of Mind, we conjecture that this arrangement of connections actually resembles an n-cube machine, within which most of the actual work is done inside distinct agencies that scarcely communicate at all with one another. Most pairs of agents neither need or are able to talk with each other because they speak such different languages. The reader might complain that this seems wrong - since a person can so easily "associate" any two ideas or states of mind, however different they seem in character. However, when we examine those associations, we often find them to be peculiarly indirect, often engaging seemingly arbitrary combinations of ideas and images. This is probably because our methods for making indirect connections must use what they can find already in our memories.

Unifying Frames and K-lines

We made substantial progress in developing and reformulating two theories that originally seemed quite separate: the new theory unifies the ideas of Frames and of K-lines. This unification leads to better ideas about how a machine could learn to formulate and use new frames that represent various kinds of knowledge.

In particular, the new theory suggests ways to coordinate knowledge representations that apply to different realms of thought – for example, to spatial, social, and linguistic concepts. By extending Patrick Winston's research at MIT on reasoning by analogy, techniques like these should lead to more effective schemes for constructing and applying analogies.

Clarifying Conceptual Dependency.

The representations proposed in the Society of Mind theory also promise to clarify and extend the representations called "conceptual dependencies" developed in the 1970s by Roger Schank's group at Yale. These were developed for combining a variety of causal, social, and mental concepts, and have been demonstrated to be effective in several areas. However, many critics have regarded them opportunistic and ad hoc, no matter that they frequently worked, because no one could see clearly why they worked. The result has been that few other researchers were willing to expend more effort on such systems . By unifying several CD concepts into a more general type of "Trans-Frame" - which also resembles the representations implied by case-grammars in linguistics, our research suggests that such systems are sounder than they at first appeared to be.

FUTURE RESEARCH

Cache-Memory and Consolidation

Most contemporary connectionist models are based on weight-learning algorithms that are rather simple and direct. In this brief section I simply want to challenge this. What happens after a person solves an interesting problem? I suspect that it is no accident that it takes a long time - typically of the order of an hour - for the records of that experience to become firmly lodged in what psychologists call long term memory. This raises two issues. First, what sort of mechanism is involved - and, second, what might be its evolutionary origin?

As for how that mechanism works, I suspect that what is involved is something like what computer technologists call cache memory. The traces of our recent mental activities are buffered in a fast-write, fast-read system, of which some, but not all, are slowly transcribed into some other, more permanent form.

As for why, one reason might simply be the time for manufacturing structural chemicals. But also that interval of time could be used for more careful filtering, involving special processes designed to select out the more important events, find efficient representations for them, attach them to appropriate retrieval cues and, most important of all, do this in ways that assign to them appropriate credits and responsibilities. Smart connectionist system may also have to be designed with similar functions in mind.

In order to obtain sophisticated learning from massively parallel machines, we have started to develop a new theory of human memory called Cache-Transfer, which incidentally may help to explain why it takes so long to make new records in human long-term memories. This scheme for short-term memory uses machinery resembling that in a computer's "cache" memory to perform several functions which include locating suitable representations in a previously unstructured memory network and then training them to serve as permanent memory. In the course of that, the system should also be able to make "credit assignment" decisions for learning, and to recruit new agents required for building more hierarchical memory systems.

Zone-Refining

The back-propagation procedure is in the class of hill-climbing procedures: each cycle computes the partial derivative of a success function with respect to each connection weight. This procedure has been shown able to lead to the spontaneous formation of significant units in the inner layers of connectionist networks. What are the practical limitations of such processes? I suspect that in order for any such process to organize a deep network, it must proceed through stages of development. First, some units located in layers near the input and the output must acquire some significance. Only then can the system proceed to develop significant units in adjacent layers. Until reached by these peripheral, growing zones of significance, the deeper intermediate layers must remain comparatively passive and stable. I have the sense of this happening in the recent experiments on parallel learning that I have examined - at least in those that yielded good results. The earliest "significant" nodes formed along short paths from input to output. Only after the formation of these sensory and motor abstractions did significant nodes develop in an intermediate layer.

Once some initial communication agents acquire some significance, we can make them available for connections to, and exploitations by, additional inner layers. The fundamental idea to represent higher level concepts in the form of nodes that act as managers for using combinations of concepts represented by lower-level nodes. This process can be repeated over and over again, introducing more and more layers over the course of time. However, it is our thesis that it will be unproductive to introduce a new layer until the previous ones have achieved some competence. (The later layers need not be inserted physically at

those times, but could be present from the start - provided that they provide signal-paths that do not vary much before the phases in which they learn.

After each new layer is established, we have an opportunity to refine and correct mistakes in the earlier layers. This is because the presence of the new, significant, inner or "hidden" units may make it possible for the first time to construct additional input classifiers and output actuators. But again, it seems to me that these refinements themselves would best be done in the course of a layer-by-layer sweep - which is why I envision the process as one of zone-refining. Without experimental experience it is hard to propose more details. Would it be better to sweep in one or both directions - or to sweep from both ends to the middle? In any case, I suggest that it is best to operate on each layer separately, while holding the others fixed. As explained in The Society of Mind there are many virtues to networks that are roughly structured into layers, because this facilitates the formation of the "level-bands" that may be vital to intelligent thought.

Whenever you employ a learning machine, you must specify a great deal more than merely the sources and destination of the data-level information. For every learning organ needs some signals to indicate what it should learn - for comparing results, testing hypotheses, and selecting suitable goals. Each choice of design must somehow determine how long the learner should persist when progress slows. How to decide when enough has been done; which particular learning procedures to employ; when to decide that things have gone badly wrong; how to determine the allocation of hardware, time, memory, and other resources? When enough such things are taken into account, our sculpturing art may look more like a weird form of management skill than anything we would recognize today as programming. For we will have to decide which agencies should provide what incentives for which others - and then we'll have to decide who will watch those watchers. As in any society, every such decision about one agency imposes additional constraints and requirements on several others - and then we have to specify how to train those other agencies as well.

Computational Linguistics

We have also developed what we call the "Copy-duplication theory of language" - a novel theory of language in which grammatical forms are treated as resources that can be exploited for expressive purposes, rather than as arbitrary forms into which communications must be forced to fit). In effect, this theory places grammar "on tap" rather than on top. It is hard to predict how long it will take to bring the new language theory into a form suitable for simulation and, then, for practical application. Some students are considering the subject, but at present we do not have suitable staff workers for a serious project of this sort. If the linguistic community's reaction to the publication of the theory is favorable enough, then such a project would be very plausible.

BRIDGES BETWEEN SYMBOLIC AND CONNECTIONIST MODELS

Research on connectionist nets will eventually make important discoveries, but those systems will remain too limited until we develop more versatile ways to control them. Symbolic representations have great flexibility because of their explicitly structured character: when you cannot find a simple way to solve a problem or puzzle, symbolic representations help in formulating networks of sub-problems and subgoals. This capability lends it self to the use of the top down strategies called heuristic programming which have developed so productively for several decades. It is hard to apply such processes to distributed representations because there is no natural way, when a problem is partitioned into smaller parts, to make corresponding partitions of the networks. How can we combine the merits of both approaches? We will try to develop systems in which connectionist networks learn in goal-directed ways, so that they can exploit what AI research has learning about goal-based reasoning. To do this, we need to develop supervisory systems that can impel the network learning processes toward producing more orderly representations. This could make it possible for higher level networks to more effectively exploit lower level ones, to form societies of networks that can exploit the knowledge embodied in them, in quasi-symbolic ways.

Most of the knowledge in a Society of Mind system is represented, not as symbolic expressions, but in the network of connections between various processes

RESEARCH TOOLS FOR SOCIETY OF MIND MODELS

Kenneth Haase has developed a test bed called NETPLAY for simulating simple Society of Mind networks. The use of this test bed should reveal unanticipated problems in localist networks and help sharpen our methods for analyzing and repairing them. These tools should also help us explore how AI's traditional explanations of mental phenomenon might be handled in the new framework. For instance, it should enable us to explore the extent to which mechanisms like frames, inheritance, and truth maintenance may be implemented in systems of entirely local interactions.

DISCOVERY PROCESSES

The processes of skill-learning are cumulative: each skill we gain is built to some extent on other skills gained earlier. We are trying to investigate the nature of such accretions by considering how a computer program might invent

new distinctions, operations, and understandings based on earlier inventions. Following Lenat's work with the AM and Eurisko programs. Kenneth Haase will continue work on a program called CYRANO that devises new representations, operations, and new domains from earlier and simpler ones.

GOALS: PROBLEM SOLVING AND MEMORY

Such terms as learning, memory, thinking, and problem solving all refer to overlapping aspects of how we change over time, with the unifying element of improvement through experience. To make such systems more competent, their memory machinery must become increasingly selective and sophisticated. This requires more than simply storing new facts, connections, or rules because large masses of memories cannot be used effectively without schemes for retrieving them in goal-related ways. It is well known that it requires a substantial amount of time for animals to convert short-term memories into permanent memories, but little is known about why that is so. We are exploring the idea that a cache-like mechanism is involved in the process of converting short-term into long-term memories, and that this serves several functions. One function is for locating suitable new representations; this is needed because different kinds of experience must be encoded in different forms. Another function is for distinguishing different roles of memories; in the course of solving a problem we learn both which operations are useful and which are deleterious - and these must be stored in different ways. Furthermore, a memory system must contribute to solving various problems of credit assignment - that is, of which operations actually affected the final results in significant ways. For all these reasons, we are working on a model in which both the construction and the later application of memories are directed by their connections with goals. This model is based on two ideas from The Society of Mind; recognizing differences by making comparisons, and representing defaults as assignments to level-bands. We expect such processes to produce memory organizations that are particularly suited for use by goal-oriented problem solving mechanisms.

Presentations 1985

Indiana University

American Psychological Association (Boston)

Massachusetts College of Art

Turin, Italy

Interview on Financial News Network

ACM Annual Technical Symposium, NBS, Gaithersburg

American Academy of Achievement, Denver, Colorado

International Arts Conference on Computers, Vancouver
NASA SMART meeting at Ames Research Center
Byte Anniversary Celebration, Boston Computer Museum
Banco Popular, San Juan, Puerto Rico
Cablenews interview
Business Week Interview
DEC Distinguished Speaker Series
Neurobiology for Neurosurgeons Conf, Woods Hole
University of Houston, Distinguished Lecture Series
CALITE 85, Melbourne, Australia
Invited lecture, University of Adelaide, Australia

Presentations 1986

University of Santa Clara, California
Boskone XXIII, National Science Fiction Conference
Yale Symposium on AI and the Human Mind
MIT Club of Washington, D.C.
Babson College
University of California at Riverside
White house conference on Crisis Management
Washington University, St. Louis, Missouri
Center for Neurobiology, Columbia University
Lecture at Univ of Brussels, Belgium
Filming for Smithsonian World TV program
Apple Computer Conference, MIT
American Academy of Achievement, Washington DC
SUNY, Southhampton
Invited Lecture, AAAI
Invited lecture, University of Geneva
Bowling Green University, Ohio
Inauguration of Media Laboratory, MIT
Computer Network Talk Issues Form (Compuserve)

Presentations 1987

Third International Conference on Thinking - Hawaii
Lecture to NYNEX corp, New York
Museum of Science, Boston
Tandem Computer, Cupertino
AAAS Youth Symposium, Chicago
Computer Institute, San Juan, Puerto Rico
IEEE Computer Society San Francisco
Nippon Electronics College, Tokyo
AI Society, Tokyo
Boston Museum of Science
BBC-TV "Horizon" NOVA-type program
NASA Conference on Telepresence
Cognitive Science Group, Harvard University
Argentine Academy of Sciences
Radio WKOX, Boston
Carpenter Center, Harvard University
WCRB Radio , Boston
IEEE, Orlando, Florida
LOTUS Development Corporation
University of Pittsburgh
American Academy of Achievement, Phoenix
CalState, Long Beach
Smithsonian World interview
Workshop on Organizational Science, MIT
Franklin Institute, Philadelphia
WNYC public radio
Congressional Clearinghouse, Washington DC
AAAI Symbolics Lecture
American Psychology Association
BBC "Antenna"
World Society Artificial Organs, Munich
Washington Post interview

Interview with Tele-university from Montreal

Boston Computer Museum

Interview in "Management Today"

IEEE Congress, Montreal

Westinghouse, Pittsburgh

"Frontiers of Science" Smithsonian talk

Naval Research Center, Washington DC

Publications

1985 "Communication with Alien Intelligence," in *Extraterrestrial: Science and Alien Intelligence*, (E. Regis, ed.) Cambridge University Press, 1985. and version in *BYTE*, April, 1985. Proposes a new philosophical theory, called the "sparseness hypothesis," about why it is possible to have apriori mathematical knowledge.

1985 "Why People Think Computers Can't," *AI Magazine*, Fall, 1982 and version in *Techology Review*, November-December, 1983.

1986 *Robotics*, editor, Anchor Press/Doubleday, Garden City, N.Y., 1986. (Book)

1986 "Are Minds Machines?" Editorial in *OMNI* magazine, September, 1986.

1987 *The Society of Mind*, Simon and Schuster, New York, 1987.

1988 "The Invention of the Confocal Scanning Microscope," to appear in *Scanning*, 1988.

1988 Prologue and Epilogue to *Perceptrons*, (with S. Papert), MIT Press, Cambridge, Mass. Enlarged edition of 1969 book.

1988 "Emotions" in book by Clynes and Panksepp (to be published).

1988 "Connectionism" in *Connectionist Models and Their Implications* editors David Waltz and Jerome Feldman, Ablex, New Jersey, 1988.

1988 "Thoughts About Artificial Intelligence," *The Age of Intelligent Machines*, MIT Press, Cambridge, Mass. (to be published).

1988 "Mind and Brain" in *Artificial Intelligence and the Human Mind* (to be published.)

Presentations

Doctor Honoris Cause - Free University of Brussels

Participants

Kenneth Haase

John Amuedo: use of multiple processes in musical analysis and pitch detection

Michael Travers

William Coderre

David Rosenthal

Alan Ruttenberg