Loosemore, R.P.W. (2013). *The Yearbook of Nanotechnology, Volume III: Nanotechnology, the Brain, and the Future*, Eds. Sean Hays, Jason Scott Robert, Clark A. Miller, and Ira Bennett. (New York, NY: Springer, October 2011)

The Complex Cognitive Systems Manifesto

Richard P. W. Loosemore
Susaro, Inc., Genoa NY 13071, USA
rloosemore@susaro.com

Abstract. In a complex system the overall behavior of the system cannot be analytically explained in terms of the underlying mechanism that causes the behavior. This paper argues that the human cognitive system is almost certainly a partial complex system, and that one consequence of this complexity is that if we try to understand human cognition by looking only for the locally-most-optimal models of all aspects of the system, we will generate models that can never converge on a unified theory. This has serious implications for the methodology of cognitive science. To solve this "complex systems problem," it is proposed that researchers move toward a new, more theoretically intensive research paradigm that shifts the focus away from local models and toward parameterized "generators" of large sets of models. These generators would then be organized using frameworks, each of which is a prototype of a unified theory of cognition, and the research methodology would involve constraint relaxation among the generated models. The paper concludes with a description of a specific framework, based on a generalized version of connectionism, and the suggestion that this new methodology can only be realized if a new class of software tools is built to support it.

INTRODUCTION

In order to understand the functioning of the brain, we need to have at least a reasonable understanding of the connection between processes and architectures described at the cognitive level and those described at the neural level. We need to know approximately how the psychology maps onto the neural hardware.

In this paper I am going to argue that there is a serious structural problem with the scientific methodology being used to understand this relationship. In spite of the seemingly rapid advances being made in neuroscience and cognitive psychology—and especially in the field of brain imaging—progress is, I would argue, much less significant than it seems because of an unrecognized technical problem in certain aspects of our research methodology.

This methodological problem—the *complex systems problem*, or *CSP*—is hard to understand and exceptionally difficult to verify. It appears that the only way to solve the complex systems problem is to make substantial changes to our research methodology, but because the magnitude of these changes is so large, there is a strong incentive for cognitive scientists to deny that the problem exists.

That is a high-level view of what the paper is about. The CSP has a number of implications, but in order to ground the discussion in something concrete, I will largely focus on how it impacts one particular aspect of cognition: the question of how behavior at the cognitive level should be explained in terms of events at the neural level.

The neural-cognitive mapping is the backbone of all cognitive science. At issue is whether the basic units of thought (concepts, symbols, etc) are implemented in the brain as single neurons, redundant clusters of neurons, distributed patterns of activation across large

networks of neurons, virtual entities with no direct connection to the hardware, or some other theoretical construct. Choosing between these rival interpretations is clearly important, if for no other reason than that data from brain scanning machines can hardly be interpreted without making assumptions about how the psychological level is related to neural events.

There are two main reasons for presenting the complex systems problem in the context of the neural-cognitive mapping. One stems from a revolution that occurred in cognitive science back in the 1980s, which turns out to have been a textbook illustration of how the complex systems problem can undermine research. That revolution was variously known as *connectionism*, *neural nets* or *parallel distributed processing*. The second reason to focus on the neural-cognitive mapping is that the connectionist revolution can be adapted to yield a solution to the complex systems problem: once we understand how the connectionist ideas were damaged by the CSP, it becomes possible to correct the damage and forge a new kind of connectionism that has the potential to overcome the CSP.

In the next section I will analyze the concept of a complex system and try to understand how the properties of complex systems might have an effect on the research methodology used by (among others) the various branches of the cognitive science community. This general account of the CSP then leads to an examination of the neural-cognitive mapping issue, especially in the form that it took during the connectionist revolution. Following that is a proposal for a generalized version of connectionism that is designed to be a partial solution to the complex systems problem. Finally, I will look at some of the broader implications of the CSP: the need for new types of software tools and the problem of overcoming the academic inertia that might stifle any resolution of the problem.

THE COMPLEX SYSTEMS PROBLEM

Systems of all kinds—whether natural or artificial—appear to come in two broad categories. On the one hand there are "regular" systems, which can be defined as those in which the components interact in ways that seem *tractable*. In the case of a regular systems we can write down equations or algorithms that define how the components of the system interact, and these equations or algorithms can be solved to derive a prediction of the system's overall behavior. Most of the systems studied by scientists over the last 300 years happen to belong in this category, because in the majority of cases we have found ways to develop a convincing, rigorous argument that takes us from a description of the system's underlying rules to a prediction of the system's overall behavior.

The other category of system is labeled "complex," and it can be defined by our inability to solve the system's underlying equations. Roughly speaking (and bear in mind that we are simplifying matters a great deal here) a complex system has an overall behavior that we can only understand by simulating the system. If we simulate the system's underlying mechanism and then observe that the simulation behavior corresponds to the original system behavior, we can say in a sense that we understand where the system's behavior comes from—but this is a very different kind of understanding than the one we get if we solve the equations and prove the behavior. If we want to get any kind of explanation for the behavior of a complex system we seem to be stuck with either a simulation or nothing: trying to solve the underlying equations will get us nowhere.

This is an extremely simplified (not to say contentious) definition of complexity, so one of my first goals in this section is to analyze the concept in sufficient detail to bring out the implications that it might have. Before getting embroiled in the detailed analysis, though, it might be worth giving a rough overview of the shape of the full argument will eventually emerge. By the time we reach the middle of the paper we should see the following sequence of steps:

- A system that is 100% complex cannot be *reverse engineered*—which means that its underlying mechanisms cannot be discovered by looking at its behavior alone. So if the universe contains any system of interest to science that happens to be 100% complex (and if we cannot directly inspect the system's underlying mechanisms), that system will be insuperably difficult to understand.
- If a system is *partially* complex, we would expect to find some aspects of the system that suffer from the same insuperable difficulty found in those that are 100% complex. This means that some features of the system's behavior will be caused by underlying mechanisms that look like they could never explain the behavior.
- Because of this difficulty, the process of finding a scientific explanation for a partially complex system will be *globally pathological*—which is to say that if our scientific methodology is driven by a search for locally-most-plausible models for every aspect of the system, we will never be able to integrate these locally-most-plausible models into a complete explanation.
- There is evidence that cognitive systems (including the human cognitive system and all human-level artificial intelligence systems) are, in fact, partial complex systems.
- Our current scientific methodology is deeply attached to the strategy of searching for locally-most-plausible models: but if cognitive systems are partially complex systems this strategy will result in an endless stream of local models that never together.

The main conclusion that emerges from this argument is that the usual divide-and-conquer approach to science will not work for cognitive systems. We would be in a position somewhat akin to that of people trying to lay square tiles on a floor that locally appears flat, but which is actually embedded in a curved non-Euclidean space. Each tiler can start working from one position and be convinced that everything is going well, only to discover that large irregular gaps appear when the separate patches of tile encounter one another. If the global nature of their problem is not understood, each person will be convinced that they can solve the problem by adjusting each pair of operations whenever a bad alignment is noticed. In normal space these pairwise efforts to improve local alignment would eventually converge on a global solution. But if the space is curvilinear, the local adjustments can go on forever without progress.

Characterizing Complexity

It might seem that the first step would be to present the most widely accepted definition for the term "complex system." This turns out to be a nontrivial task, because even complex systems scientists have not been able to reach a consensus definition (Mitchell, 2008). In fact, some critics of the field (Horgan, 1995) have used this state of confusion to argue that no proper definition will ever emerge.

In order to get past this definitional difficulty it is necessary to find a way of defining the term that includes an explanation of why it should appear to be such a fractured idea at the

moment. Reconciliation between the competing interpretations of the concept can best be achieved by clarifying why it is that there is so much competition and disagreement.

Accordingly, I will now try to develop an extended account of what complex systems are, together with an explanation for why we currently have several conflicting interpretations. Then, having laid the groundwork, we can move on to ask how this idea might have on the scientific methodologies we use to study such systems.

Systems

A *system* consists of a number of *components*, and these components engage in certain *interactions* with one another.

Every system exhibits an overall *behavior* that is a result of the interactions between its components. Viewed from the outside, the behavior of the system is what we notice first, whereas the components and/or interactions might initially be hidden. From a cause-effect point of view the behavior is an effect, while the components and interactions are the cause.

It is often convenient to use the term *mechanism* to stand for a combination of the components and their interactions. In that case, we would say that the (observable) behavior of a system is a consequence of the (often hidden) mechanism that underlies the behavior.

Strictly speaking there are two features of a system that are caused by the mechanism: the behavior, and the *form* (or shape, or structure) of the system. The first is a dynamic feature, the second is static. For the sake of narrative convenience the term *behavior* will often be used to signify both of these. So "behavior" can be either a dynamic or static feature of the system.

Regularities

When we talk about a system's behavior what we really mean is a *regularity* in the behavior. These two are not the same, because a given system can have many different regularities in its behavior, and these can exist at many levels of description. The behavior of hurricanes, for example, includes one regularity that is the spiral shape, but there are other regularities such as the role played by hurricanes in the world's ecosystem, or the typical regions in which they occur, or the dynamical development of a single hurricane.

There is no reason why all of the different regularities that we might see in a system have to be of the same type, or follow the same rules. In fact, the concept of a regularity is observer-dependent and often quite subtle, so it is not very meaningful to talk about "all" of the regularities possessed by a given system. A regularity is a construct that we see in the behavior.

This distinction between system, behavior and regularity is important, but for convenience we often blur the distinction by using the words "system" or "behavior" to describe one particular regularity. We might say, for example, that Newton used his inverse square law of gravitation to explain the motion of planets in the solar system—but what we really mean is that he explained a certain cluster of regularities in the behavior of the solar system (namely, Kepler's Laws). Other kinds of regularity, like the pattern of temperatures on the surface of the planets, were not addressed by his theory.

A regularity is nothing more than a non-random *pattern* in the behavior of a system. The concept of a pattern is quite vague, so regularities are not always captured in concise laws

like those discovered by Kepler. In some cases we might observe a pattern in a system's behavior but find it hard to write down an objective, closed-form description of the pattern. In spite of this, though, elusive regularities can still demand that we give them a scientific explanation.

Explanation

Explaining a regularity entails much more than just finding the correct underlying mechanism. Before Newton finished work on his law of gravitation some of his contemporaries had already suspected that there was an inverse-square force of attraction between the planets and the sun. But at that point this was just a candidate mechanism, because nobody could prove that this mechanism led unambiguously to a prediction that the orbits would follow Kepler's laws.

At the risk of laboring a point that is surely second nature to any scientist, the process of finding an explanation involves two steps, in which a candidate mechanism is first generated (the hypothesis), and then the candidate is used to construct a chain of inference that leads to a prediction of the behavior. We first go "backward" from behavior to candidate mechanism (the conceptual brainstorming that Newton and others did before they guessed that there might be an inverse-square attractive force), and then we turn around and go "forward" from candidate mechanism to behavior (which in Newton's case involved the invention of the calculus, so he could solve the inverse-square force equation).

The reason for stating the obvious here is that this backward step from behavior to candidate mechanism is significant, and often underestimated. Folk wisdom portrays the art of scientific discovery as an inspiration-plus-perspiration effort, in which the initial inspiration is a blinding flash of insight that enables the scientist to come up with the (in hindsight, correct) hypothesis. Then comes the perspiration phase when the implications of the hypothesis are rigorously elaborated to show that the mechanism does lead to the observed behavior. But by shrouding that first step in the concept of "inspiration" we do a disservice to the very concrete cognitive processes at work when a hypothesis is created.

We know little about this backward pass, from observation to hypothesis, but it seems safe to say that—taking Isaac Newton as a prototype once again—many features of the behavior of planets, moons and apples contributed to a chain of (mostly unconscious) clues that pointed toward the idea that objects falling on earth were connected to planetary orbits. Newton came up with a candidate mechanism that explained Kepler's laws not by magic, luck or blind guesswork, but by being sensitive to many factors that pointed toward the correct mechanism.

A Break in the Forward Path

One surprising feature of the systems studied by scientists over the last 300 years is that in the overwhelming majority of cases there exists a rigorous chain of inference that goes from the candidate mechanism to the predicted behavior.

This may seem a trivial observation, but it is only trivial if we assume that explanations are always there to be found. There is really nothing necessary about the existence of such proofs: it is an empirically interesting fact about the universe that so many of the systems of interest to science turn out to have concise, provable connections from candidate mechanism to behavior. There is no reason why this should always be the case: there is nothing in the

structure of the universe that guarantees that every mechanism is connected by a clean proof to the behavior it gives rise to.

In particular, there is no reason to assume that *if* a regularity exists in the behavior of some system, *therefore* a deducible connection from the mechanism to the regularity can be found. Naive intuition might lead us to suppose that if a system behaves in some elegant, structured way, this is in itself an indication that somewhere beneath the surface there exists an elegant explanation for that behavior. But however compelling this might seem—and however frequently it might have happened that way in the history of science—there is nothing logically necessary about the existence of a compact explanation, given the existence of a regularity.

Are there any examples of systems where a behavioral regularity exists, but no explanation can ever be had? This is a problematic question: we could never know that *no* explanation will ever be possible. We might suspect a system of being beyond explanation in this way, but there is always the possibility of being surprised, tomorrow, by an unexpectedly new and elegant proof.

What we can say is this. To an omniscient scientist, with access to all the knowledge that could possibly exist, it might be knowable that there really are systems in this category. But with our limitations we can only say that we *suspect* some systems of having no explanation that connects the underlying mechanism to an observable behavior. As a first approximation, then, the definition of a complex system is that it *appears* to belong in this category.

If all of a system's behavior regularities appear to have no explanation, then we can say that the system is 100% complex. If some regularities are explicable while others seem beyond explanation, then the system is *partially complex*.

Numerous examples could be cited, but Stephen Wolfram (2002) has investigated a notably extensive set. Wolfram (2002), in fact, uses the term *computationally irreducible*, as an alternative way to refer to complex systems. This means that we cannot compute the behavior of the system by using an algorithm or equation that is more compact (more reduced) than the algorithm or equation that is encapsulated in the system's mechanism itself.

Notice that so far this definition only references the forward path of the explanation cycle: a system is complex if the route from mechanism to behavior is broken. This begs an interesting question that we now consider.

A Break in the Backward Path

Suppose that a system is complex in the above sense, so there is no way for us to extract a prediction about its behavior given knowledge of its mechanism. Would it nevertheless be possible for some human (or machine) genius to look at the behavior and traverse the backward path from behavior to mechanism? Could someone intuit the correct mechanism that was behind a set of behavioral observations, even though they would never be able to develop a proof that their hypothesis was correct?

If this were possible it would significantly lessen the impact of complex systems. After intuiting the correct underlying mechanism, the scientist could then feed this into a computer simulation and use the simulation to prove that the candidate mechanism was valid. There would be no need for a mathematical proof or argument to go from mechanism to behavior.

It is difficult to collect information about whether this ever happens, because in practice the known examples of regular systems and complex systems tend to be treated differently. The regular systems that have dominated most of our science have always been subjected to that backward pass (for the obvious reason that this is an indispensable part of building an explanation). But in the case of many of the complex systems that have been studied, we have *invented* the mechanisms that define the system, so we have almost never tried to work backwards from known behavior to unknown mechanism.

We can go one step further and note that in those cases where a natural complex system has been studied, there are always some aspects of the system that are regular, so when the underlying mechanisms are not obvious, and have to be discovered, the discovery was initially done without using the complex aspects of the system. Having thus uncovered the mechanism by doing normal science on a (largely) regular system, the complex aspects of that system could then be studied in the same way as with artificially constructed complex systems—namely by exploring the consequences of the mechanism using simulations.

Given this observation, and the fact that there are no known examples (at least, known to this author) of artificial complex systems whose behavior was written down first, then used to work backward to the mechanism that gave rise to the behavior, we can make the following conjecture:

- If a system does not have a logico-mathematical path leading from mechanism to behavior (if there is no "forward path"), then the absence of this path means that the backward path (from behavior down to mechanism) cannot be traversed either.

What this conjecture says, in effect, is that when a scientist first encounters some observable behavior that needs to be explained, the process of generating a viable hypothesis about the cause of the behavior (a viable candidate mechanism) will only be feasible if there exists a proof or argument that leads in the other direction, from hypothesis to behavior. If the system is such that the only way to get from candidate mechanism to behavior is via a computer simulation of the mechanism, then the subtle cognitive apparatus that scientists use to come up with a hypothesis about the system simply cannot operate. The process of scientific discovery cannot happen unless there is a non-simulation route that can be used to explain the behavior of the system.

Some obvious caveats need to be mentioned. If the system is simple enough it is possible that a simulation can be done in the head of a human scientist, and in that case the conjecture would not apply. Also, it would be feasible to work backward from behavior to mechanism in those cases where the number of possible mechanisms was sufficiently small that we could mount an exhaustive search through all the possible simulations.

This feature of complex systems is not something that gets a great deal of attention because, as I explained above, people do not usually try to invent complex systems that have a particular behavior. This activity could be described as *reverse engineering* a complex system, and as a general rule it is simply assumed by complex systems researchers as being too obviously infeasible to be worth considering. The definition of complexity, after all, is that the behavior is emergent and therefore not what would have been expected from the mechanism—and this unexpectedness is normally assumed to imply that reverse engineering is the one thing that cannot be done.

Recipe for Complexity

When using the terms "complex system" and "complexity" we need to be clear about whether we are referring to observable features (the behavior) of systems, or to some structural characteristic (aspects of its mechanism) that might be giving rise to those features.

So, if we look at the empirically observed characteristics of known complex systems, it is possible to give a list of design ingredients that tend to make a system complex:

- The system contains large numbers of similar computational elements.
- Simple rules govern the interactions between elements.
- There is a significant degree of nonlinearity in the element interactions.
- There is adaptation (sensitivity to history) on the part of the elements.
- There is sensitivity to an external environment.

When the above features are present and the system parameters are chosen so that activity does not go into a locked-up state or an infinite loop, then there is a high probability (though by no means a certainty) that the system will show signs of complexity.

Notice the phrase "signs of complexity." It is all too easy to employ words such as this when we are really trying to signify the *disconnection* between the mechanism and behavior (the lack of a rigorous explanatory path that leads from one to the other). What is confusing is that the phrasing seems to imply that the behavior *per se*, or the mechanisms *per se*, are showing some signs of being "complex". In truth, it is only the disconnect between them that makes the system complex. It is often convenient to speak loosely about the behavior or mechanism being "complex," when all that is meant is that the two are disconnected.

Absence of a Clear Diagnostic

One fact about complex systems is especially subtle:

- It is (virtually) impossible to find a compact diagnostic test that can be used to separate complex from non-complex systems, because the property of "being a complex system" is itself one of those behavioral regularities that, if the system is complex, cannot be derived analytically from the low-level mechanism of the system.

The "virtually" qualifier, above, refers to the fact that complex systems are not completely excluded from having behavioral features that are derivable from local mechanisms. So it is conceivable in principle that a system could have no explanation for a significant chunk of its behavior, while at the same time this lack of existence of an explanation could itself be a provable fact about the system. While conceivable, however, this would be a bizarre situation—the proof would have to be rigorous in spite of the fact that it contained a concrete reference to a thing (the unexplainable regularities in the behavior) that could not be connected to the rest of the facts about the universe through any kind of formal structure or proof. It is hard to see how any proof could still count as a proof, while containing such an intangible.

This is an interesting result, because it means that complex systems are defined in such a way that the whole concept can only be coherent and internally consistent if the definition never becomes precise. One consequence is that when we debate whether a particular class of systems (e.g. intelligent systems) might be complex, the debate cannot include a demand for a definitive proof or test of complexity, because there is no such thing.

We can now begin to get some traction on the problem mentioned earlier, that different researchers define complexity in different ways. If anyone did produce a perfect, closed-form definition of what complexity was, that definition would auto-destruct, so perhaps this impossibility of finding a perfect definition is having an effect on all attempts to build comprehensive definitions. Although this does not explain all of the variance to be seen across the different efforts to pin down the nature of complexity, it does shed some light on at least one source of confusion.

Deniability

The fact that a complete definition of complexity is impossible leads to another consequence that has far reaching implications. If it should ever happen that complexity effects become a nuisance to some scientific community—for example, if those effects seem to imply that the community should adopt a radical change of methodology—the easiest strategy for the community to adopt is to deny the existence of the effects altogether. This is easy because of the extraordinary difficulty of defining what is and is not a complex effect—and therefore what is or is not a *consequence* of complexity. It is always possible for the skeptic to insist that concrete proof be given that complexity effects are responsible for some situation. Then, in the absence of such a proof, the situation can instead be blamed on a mere difficulty with the understanding of a regular system.

These circumstances have already (ostensibly) arisen in economics and elsewhere (Waldrop, 1992). Substantial conflicts have taken place between groups promoting the opposing viewpoints—for or against the importance of complexity—and it is arguable that these conflicts have been exacerbated by the difficulty of distinguishing complexity from regular system effects that have just not been fully analyzed yet. If the above analysis is correct, and the indefinability of complexity is intrinsic to its nature, then the battle between these opposing viewpoints may be even more protracted than it usually is in a scientific paradigm conflict.

Partial Complexity Versus Full Complexity

Most systems that are complex at all, are only *partially complex*: some aspects of their behavior can be understood as a regular consequence of some mechanism, while other aspects seem emergent.

This partial complexity can manifest in many forms. One of these is that a system can have several levels of description, and some levels can be complex while others are regular. One example of a multilevel system is the well-known cellular automaton invented by J. H. Conway, known as Game of Life (Gardner, 1970). In this system there are some very simple rules that determine whether each cell of a square grid is in the *on* or *off* state, at every cycle of a global clock. Certain patterns of initially-on cells will result in cyclic activity: the pattern repeats after a fixed number of clock cycles. There are many known patterns that have periodic behavior in this way, and from our point of view the behavioral regularities of interest would be the shape of the stable patterns and the period of each one. As far as we know, there are no forms of analysis that would allow us to input the rules of Game of Life and receive, as output, a prediction of the shape and period of all the stable creatures that can be found in this system.

At the level of the first batch of creatures to be discovered, the regularities are complex, because of their lack of derivability from the mechanism. But it is possible to use some of these basic patterns as ingredients for higher-order patterns, and those higher patterns can be constructed in such a way that they perform quite predictable, regular-system behaviors. Indeed, it has been shown that the patterns can be arranged in such a way that a complete Turing Machine is built inside the system.

Other ways to encounter partial complexity are more straightforward. A system can have several components, governed by different mechanisms, with only some of these being complex. Or, a number of different regularities can be due to the same mechanism, but with differences in their complexity. Gravitational motion in the solar system, for example, approximates very well to a regular system, but only if we ignore such effects as Pluto's occasional bursts of chaotic behavior, and the braiding patterns observed in planetary ring systems due to systematic influence of nearby moons.

Some partially complex systems can be *primarily regular*, while others can be *dominated by complexity*. The solar system would be primarily regular, because there are dramatic regularities in the orbits of the planets that enabled Newton to derive an extremely accurate account of the underlying mechanism. Whenever a system has enough regular aspects to it that we can use those regular aspects to work backward and uncover all of the underlying mechanism, we can categorize the system as primarily regular. If, on the other hand, there are significant behaviors that seem complex, and we do not have any easy way to go around to the back door, so to speak, and use regular aspects of the system to uncover the mechanisms, we would classify the system as dominated by complexity, or as containing significant amounts of complexity.

Are Cognitive Systems Complex?

How do we decide whether intelligent systems should be treated as containing significant amounts of complexity? There are some aspects of human intelligence that seem to involve sequences of logical inference that are governed by rules, so from that point of view the system looks regular. And there are plenty of other regularities to be found, across all the paradigms of experimental cognitive psychology.

But of perhaps greater significance is the fact that the core engine of our intelligence—the mechanism that creates, develops and deploys *concepts*—is known to involve a host of subtle interactions and sensitivities. Concept construction and deployment is one of the most poorly modeled of all aspects of cognition, in the sense that we are still grasping for the correct metaphors with which to characterize them (are they prototypes?, exemplar-driven?, distributed patterns of microfeatures?), we still have many choices of theory to describe their developmental aspects, and it is still not possible to build working artificial intelligence systems that construct and maintain concepts of arbitrary levels of abstraction, using only raw real-world input.

The creation and development of concepts is the place where we would most expect to see complexity, because this is where we have evidence of *intractable interactions* between the components of the system. We can combine concepts in a seemingly infinite variety of ways, and we can use them with degrees of flexibility that appear to be unbounded. Almost every time we use a concept we adapt it to the specific context in which it is used. All of this

flexibility, context-dependence and combinability seems to point to a system in which component interactions are out of control.

Without pushing the case as far as it might be pushed—by listing a full catalogue of examples that seem to indicate complexity—let's step back for a moment and consider what the goal is here. Are we trying to decide whether there is conclusive evidence that a significant amount of complexity is present in human cognition? If this were the goal, it would be a risky one: we have already seen that it is impossible, in principle, to prove that a system is complex. It seems that if we had the lesser goal of showing that *there is a substantial risk that complexity is present*, we might be able to close the case immediately. I submit that this is already done, and is widely accepted by the cognitive science community. The features of concept building described above have been remarked upon throughout the history of the subject—so much so that it is almost a standing joke that when anyone tries to pin down the meaning of a concept in an algorithmically closed form, someone else will immediately produce a counterexample.

Speaking informally, cognitive scientists seem quite ready to concede that many aspects of cognition (including concept mechanisms) show evidence of complexity. They may not want to take the next step and admit that this has great significance, but it is enough for our purposes to note that complexity is widely accepted to be present. And perhaps present to a significant degree.

The Risk of Complexity

One of the main goals of this paper is to argue that complexity has much greater, damaging consequences than has been appreciated. Ideally, this would then be accompanied by a proof that cognitive systems must be complex systems, and we could then draw the obvious conclusion that these consequences will have an impact on cognitive science. But since we cannot, in principle, give a proof that cognitive systems contain significant amounts of complexity, the best we can do is show that there is a substantial risk that they do. In view of that risk, we would then need to take action.

I believe that the risk of complexity that is indicated by the known properties of the concept mechanism is more than enough to establish that we need to ask whether the consequences of complexity would really be that severe. It is that last question that we now consider.

COMPLEXITY AND SCIENTIFIC METHODOLOGY

Given all of the preceding arguments about the definition and characteristics of complex systems, what can we conclude about the way we study systems that appear to be complex? In particular, what impact might this have on the methodology of the cognitive sciences?

If cognitive systems are partial complex systems, then what this means is that there are some behavioral regularities that can only be explained by mechanisms that do not look as though they would ever explain those regularities. This is just another way of saying that the link from mechanism to behavior is broken in those cases—broken in both the forward (proof) phase and the backward (hypothesis generation) phase. If we were to somehow discover what the true mechanism was, for one of those regularities, we would look at the mechanism, look at the behavior, and wonder how the one could ever have been expected to give rise to the other. We might then do a simulation of the mechanism, note with satisfaction that it

yielded the same behavior, and know for sure that this was the correct mechanism, but other than that we might never be able to say where the behavior came from.

Notice, though, that the absence of a backward path would make it extremely difficult (perhaps impossible, for all practical purposes) to ever deduce what the mechanism was. If (as was conjectured earlier) the process of coming up with viable hypotheses or candidate mechanisms is dependent on the existence of a regular system, that process might not be feasible.

Putting all these ideas together, what all of this means from a practical point of view is that if we approach the scientific analysis of a partial complex system by looking for models of local aspects of the system that are always rational and regular—there is always an understandable relationship between the behavior being explained and the model being used to explain it—then will always be setting ourselves up for failure on some of those models. The system cannot be entirely made from components that have a non-complex relationship to the behaviors they produce. By assumption the system is partially complex so it must be the case that at least some of those models must have a pathological relationship to the behaviors they generate. We may never know which components are to be expected to have this pathology (although we can sometimes make a shrewd guess), but we can be sure that if there is some reason to suppose that the system as a whole is partially complex, then somewhere there will be trouble.

This innocuous-looking point has profound consequences. In a truly fundamental way our science is built on the idea that we can understand the world by using occam's razor to find the simplest, most elegant explanation for all the components of a system, then combining these separate understandings into a unified understanding of the system as a whole. But in the case of an egregiously complex system this is a disastrous strategy: it will always miss the truth.

What would happen if, in spite of the danger, we did go full steam ahead and apply the usual scientific strategy? The naive conclusion might be that in the case of those system components that deserved a complex explanation, we would see our models breaking and eventually conclude that this component needed to be treated differently. I think that is probably too clean. More likely, we would find that we can always build models of all components of the system, but some of those models will just be locally applicable, or will reference such minute and insignificant aspects of the system that they are actually avoiding the features that are complex. In other words, local model building will not fail, it will just produce an unending stream of poor quality models.

And as this model building continues, two other processes will be observed. One is that each model will be extendable only at the cost of excessive complications. In order to make the model more general or apply it to more cases, it will have to be extended or elaborated with arbitrary extensions that eventually turn it into a theoretical kludge. The second process that will be seen is that when two models collide, the process of integration (which means, adaptation of each to make a unified whole) will come only at great cost: again, the result will be excessive complication. More likely than the combination of models, though, would be their insularity: researchers in different paradigms will simply decline to integrate their model with others.

All of this can be expected to happen as a result of a situation in which the complexity of the system was being denied or not acknowledged. Local model building would work toward

the goal of a complete explanation for the system that was complexity-free, but since no such non-complex, complete explanation exists, this goal would be unattainable, and the separate model-building efforts will only become more complicated an less plausible as time goes on.

This is, arguably, exactly what is happening in the cognitive sciences, both within disciplines and across them.

Solving the Problem

Suppose that we were faced with a system whose functioning we wish to understand, and which we have reason to believe could contain significant amounts of complexity. We know that choosing the locally most optimal models of the various aspects of the system is a poor strategy. What strategy should we use instead?

The simple answer must be to stop focusing so narrowly on locally optimal models. What does that mean, in practice? It can only mean that we have to choose a wide variety of models for each component of the system that we wish to model, where previously we might have chosen only one. Somehow we must test the viability of all these models in parallel.

Inventing single models is hard enough: inventing large sets of models to explain just one set of observations is even harder. This is a process that cries out for some kind of automation: somehow we must start to think of models not as individual hand-crafted works of art, but as entities that can be described in terms of parameters. Then, with this new vision of models as sets of choices for the parameters—these being *design* parameters, of course—we can start to produce large sets of models, with each model being one point in a multidimensional space of parameters.

This would entail a radical change of mindset, from models to *parameterized generators* of models. Instead of thinking of models as separately interesting things, we have to become interested in clusters of model-design parameters. This would clearly bring a significant change to cognitive science. There would still be room for cognitive scientists to interpret the results of an experiment by conceiving a new model, but such an act of model creation should then lead to a mental disassembly of the model, to see whether it can be built with already existing parameters, or if new parameter-concepts need to be invented to capture it.

Two features of this parameter-focused methodology are worth noting. One is that researchers need to keep themselves agnostic or impartial about the merits of individual choices of model. Yes, a particular model may look as though it elegantly explains some phenomenon, but this elegance may turn out to be ephemeral. Why be so dismissive about good-looking models? Because the narrower the domain of application of a model (the more it is specially designed to explain just one experimental paradigm) the more likely it is to work. The perfection of fit to its data may mean nothing more than that the model was so unconstrained by other aspects of cognition that there was plenty of room to tailor it to work in its narrow domain. We need to be less impressed by such models.

The second aspect of the parameter-hunting methodology that I want to draw attention to is that we should be actively looking for models that do *not* look as though they would be likely to explain the data. So, rather than try to limit the choice of model-design parameters so that as far as possible the models picked out are likely-seeming candidates, we should be actively looking for systems (models) where we would be surprised if the correct overall behavior was to emerge. We should be looking, in other words, for models that are complex.

Basic Models, Glue Models and Architecture Models

The models we have talked about so far are (by default) those designed to explain the data coming from specific experimental paradigms at the cognitive psychology level. These are the front lines of model-making. We can call them basic models, to distinguish them from two other general categories of model that will now be described.

When we explain various aspects of cognition we normally focus on particular experimental data, but hiding in the background there are always some implicit ideas about how these separate models might integrate with one another. How they would connect in the mundane sense of exchanging data. Although these other ideas are often ill-formed or implicit, they will eventually need to be made more explicit, and as they become explicit they effectively become models in their own right. To distinguish them from the basic models we can refer to them as *glue models*. So, for example, a glue model would be needed to explain how the cohort model of spoken word recognition (Marlene-Wilson and Tyler, 1980) can interface with the Bruce and Young (1986) model of face recognition ... and so on for all other pairs of models that might conceivably be connected.

In addition to the glue models (though overlapping them somewhat), there would also be a need for a hierarchy of *architecture models*, which would serve to integrate and unify the basic models and glue models. One kind of architecture model might be symbol-based and handle language comprehension. Another might describe the aspects of cognition that involve object understanding.

Frameworks and Constraint Relaxation

The picture that is now emerging is of large numbers of parameterized model-generators, some of them specific to distinct experimental paradigms, some acting as glue between the basic generators, and some supplying a hierarchy of architecture models. All of these generators are sources of variation because each leads to many model choices.

With these generators in place, the process of explaining human cognition is then about doing constraint satisfaction on these populations of generators. The models that emerge from the generators are capable of constraining one another in various ways, so the challenge is to discover and express these constraints, then explore the space of solutions looking for a minimization in the breakage of constraints.

Taking one step back for a moment: if all of these points of variation were to be seen as a single, homogeneous landscape that had to be explored, with no structure and no restriction in the combinatorial possibilities, the methodology might look somewhat daunting. As described, it makes cognitive science look like nothing more than single numerical computing project with billions of parameters. In fact, however, the effort is likely to be organized around some separate overall paradigms that we can call *frameworks*. A framework is just a prototype of a unified theory of cognition: a set of ideas that loosely organizes a particular set of choices for the model generators. Each framework is a more or less coherent vision of the architecture of cognition, but it would not by any means try to include all the possible models that could be conceived. In that respect it is limited: it carves off one (hopefully coherent) chunk of the space of all possible cognitive systems, so that exploration of models within that part of the space can be treated as one paradigm.

CONNECTIONISM AND THE MOLECULAR FRAMEWORK

Having set out what seems to be a reasonable first step toward a solution of the complex systems problem, in this section I want to achieve two goals. Both involve the set of ideas about cognition known as *connectionism*, or *parallel distributed processing*. The first goal is to use connectionism as an illustration of how the complex systems problem can do damage to a research paradigm.

The second goal is to take the connectionist ideas, repair the damage done by the CSP, and then present a new framework (in the sense described in the previous section) that can be seen as a generalized form of connectionism. By describing this molecular framework here I hope to give a slightly more concrete illustration of the direction that the proposed new methodology might lead us.

Connectionism and Constraints

When connectionist ideas first came to prominence in the 1980s, the core principle was often presented as being about the use of neuron-like processing units. Although this was a big part of the story, the background motivation was actually more general than this: it was about finding ways to build cognitive models in which *multiple simultaneous constraint relaxation* was the single most important feature. McClelland, Rinehart and Hinton (1986) gave a catalogue of examples in which cognition seems to involve mutual simultaneous constraints:

- Reaching and grasping
- The mutual influence of syntax and semantics
- Simultaneous mutual constraints in word recognition
- Understanding through the interplay of multiple sources of knowledge
- Stereoscopic depth perception
- Perceptual completion of familiar patterns
- Content addressability of memory

The significance of these, and other, aspects of cognition is that they seemed to point toward a type of model that involved constraints. The challenge was to find ways to build such models, and what was inspiring about the connectionist revolution was that a group of carefully designed algorithms—Boltzmann machines, interactive activation, back-propagation, etc.—were given as concrete examples of what could be done with networks of simple processing units that weakly constrained one another's state.

Interestingly, though, as the connectionist movement matured it started to restrict itself to the study of networks of neurally inspired units with mathematically tractable properties. Network models such as the Boltzmann machine (Ackley, Hinton and Sejnowski, 1985) and backpropagation learning (Rumelhart, Hinton and Williams, 1986) were designed in such a way that mathematical analysis was capable of describing the global behavior—the primary characteristics of these systems were not complex.

The problem with this emphasis on non-complex models is that, if the CSP is valid, this reliance on mathematical tractability would restrict the scope of the field to a very small part of the space of possible systems. The original connectionist researchers could have taken their original inspiration and used it to explore vast numbers of systems in which weak

constraints were operating, but instead they tried to fix the known weaknesses of the early models by either looking for better (but still mathematically tractable) models, or by combining the known models into hybrid architectures.

The subsequent history of connectionism was disappointing to some. After great initial promise the field tended toward stagnation, with no dramatic solutions to the larger problems of cognition, to match the bold solutions to some of the simpler problems, which all happened within a few years of one another at the start of the revolution.

It seems that the field was driven by one implicit assumption: that the best (and perhaps only) place to look for models in which constraint satisfaction was the driving force, were those models that could *provably* be shown to have the correct type of behavior. From the point of view of the complex systems problem, this aversion to complex systems was a grave mistake.

A Molecular Framework for Cognition

If we wanted to explore models of cognition in which mutual simultaneous constraints played the largest possible role in the functioning of the system, but if we also did not wish to make the connectionist mistake of forcing our models to be non-complex, what kind of framework might we use?

There are many possibilities, of course. What I want to do now, though, is describe one of these possible frameworks in a little detail, both as an example of where connectionism might have gone if it had not restricted itself, and as an illustration of one direction that we might go next, now that we understand the force of the complex systems problem

This "molecular" framework can be seen as using one simple idea as its point of departure. When the relative strengths and weaknesses of connectionism were first discussed, many of the weaknesses could be traced back to the fact that the units that did the constraint relaxation were locked into fixed positions within the system. Since people imagined them as neurons, they were fixed in place as real neurons were. It often seemed as though cutting the strings that tied the units down might lead to a resolution of these problems... but cutting them loose would also mean that the mathematical formalism that guaranteed their performance (the backpropagation algorithm, et al) would be lost.

Since we do not care about those algorithmic guarantees, why not imagine a type of object that constrains other objects, but which is free to roam around (like some kind of atom) in a space? As the atom roams around it forms temporary bonds with other atoms, because this is the only way to preserve the idea that the atoms constrain one another. This starts to look like a picture in which transient, ephemeral molecules are being made and unmade in a space.

Can a cognitive framework be built around this generalized version of connectionism? The remainder of this section is a very brief sketch of how such a framework might look.

Atoms and Elements

At the heart of this framework there is a distinction between instance and generic versions of the concepts stored in the system.

In much of cognitive science the idea of a concept is used as if there were only one entity encoding the concept. So, for example, theorists will talk about *the* [coin] node becoming strongly activated, or about the priming effect this can have on *the* [bank] node. It is tempting

to imagine a large network of nodes (or even neurons), with a [coin] node and a [bank] somewhere in the network, and with vast numbers of connections between all the nodes.

But any complete model of a cognitive system must include instance nodes that represent the particular entities involved in our thoughts at a given moment: nodes that represent, not coin in general, but the particular instance of the word coin that is being witnessed right now. Any realistic model of cognition must make explicit allowance for these instances, and it turns out that this can have a drastic effect on our theorizing. Instance nodes do not sit quietly in a fixed network; they are created on the fly, they have a relatively short lifetime, and the connections between them are extremely volatile. Furthermore, a complete model should explain how the generic concepts are built up from repeated exposure to specific instances.

The primacy of instances is the core concept behind the proposed molecular framework. There is a deep assumption that, in practical terms, the place where these instances are created, interact, and have their effects on the rest of the system is likely to be far more important than the passive network of generic concepts.

In other respects, the framework is little more than a conjunction and distillation of the most common features of many local theories, though with a bias toward the abstract motivations that drove, among others, McClelland and Rumelhart (McClelland et al. 1986).

Foreground and Background

In our framework there is one main type of object, and two main places.

The objects are called atoms, and their main purpose is to encode the smallest packets of knowledge (concept, symbol, node, etc.). Atoms come in two sorts: generics and instances. For each concept, there is only one generic atom, but there can be many instance atoms. In what follows, the term atom on its own will usually be understood to mean an instance atom.

The two main "places" in this framework are the foreground and the background.

The foreground roughly corresponds to working memory, and is the place where instance atoms are to be found. The foreground is an extremely dynamic place: Atoms are continually being created, and while they are active they can move around and form rapidly changing bonds with one another. The sum total of all the atoms in the foreground, together with the bonds between them, constitute what the system is currently thinking about, or aware of.

The background is approximately equivalent to long-term memory, and is just a store of all the generic atoms from which instance atoms can be made. When an instance atom is in the foreground, it maintains a link back to its generic parent in the background. The background is more or less passive; the foreground is where everything happens.

Note that atoms do not necessarily encode concepts that have names. Some of them capture regularities at a subcognitive level, and for this reason the foreground contains some activity that the system is not routinely aware of, or that it does not find easy to introspect or report on (see Harley 1998, for more detail on this point).

Active Representations and Constraints

So far, this is all sufficiently general that it could be the outline of many different theories of how the cognitive system is structured. But now we will make a commitment that distinguishes this framework from many others: The representations in the foreground are not passive tokens of the sort that are meant to be assembled and used by some external

mechanisms, they are active representations. In other words, although the atoms encode knowledge about the world in just the way you might expect, they also encapsulate a set of mechanisms that implicitly define how this knowledge is used by the system.

How do the foreground atoms do this? Broadly speaking, each atom contains (and continually updates) a set of constraints that it would like to see satisfied by its neighbors in the foreground. For example, the [chair] atom would prefer to see a group of atoms around it that encode the characteristics and components of a typical chair, and these preferences, encoded inside the atom, are what we refer to as the constraints it is seeking to satisfy.

An atom will not just passively seek a place where its constraints are satisfied, it will actively try to force its neighbors to comply with its constraints. Its behavior is a mixture of "Do my neighbors suit me?" and "Can I change my neighbors to better suit me?" An atom can engage in several kinds of activity in pursuit of its goals: It can try to activate new atoms that it would like to see in its neighborhood, or deactivate others that it does not want to see, or change its internal state, change the connections it makes, and so on.

Not all of the atoms in the foreground are successful in their attempts to satisfy their constraints. Many get woken up because something thinks they might be relevant, but after doing their best to fit, they fail to establish strong bonds and become deactivated. A substantial number of atoms, however, do find themselves able to connect to an existing structure and thereafter play a role in how that structure develops.

As a general rule, the constraints inside an atom are supposed to encode a "regularity" about the world—say, the regularity that if something is a chair, it must possess constituent parts that could be construed as legs.

Relaxation

The process in which an atom looks around and tries to take actions to satisfy some local constraints can be characterized as relaxation. Everything that happens in the foreground is driven by relaxation. If a change occurs somewhere in the foreground, atoms adjacent to the change will try to accommodate to it, and this accommodation will propagate out from the starting point like a wave, or domino effect.

There does not have to be only one type of relaxation happening at every place; several such processes can occur simultaneously, originating in different parts of the foreground. Almost everything that happens in the foreground can be attributed to changes that occur in a particular place and then send waves of relaxation that propagate across the foreground. Strictly speaking, this is dynamic relaxation, because the foreground never settles into a quiescent state where everything is satisfied; it is always on the boil.

The Foreground Periphery

At the edge of the foreground are a number of areas that connect to structures outside the foreground. One part of the periphery has input lines coming from sensory receptors, while another has output lines that go to motor output effectors. These two patches on the edge of the foreground are responsible for much of the activity among the atoms. When a signal arrives from a low-level sensory detector (or more likely a data-driven system that preprocesses some raw input signals), the result is that one or more atoms are automatically attached to the foreground periphery to represent the signal. These initial atoms then have an effect on nearby atoms.

Another important part of the periphery is a connection to mechanisms that govern the goals, drives, and motivations the system is trying to satisfy. These can be thought of as lying "underneath" the foreground, in the sense that they are more primitive than the cognitive work that goes on in the foreground proper. The work of these motivational/drive systems is not simple, but the effect of their actions is to bias the activity in one direction or another.

Sources of Action

So far, atoms have been characterized as if their role is just to encode knowledge, but in fact, representing the world is only part of what they can do: Some atoms initiate actions, and some may encode mixtures of action and representation.

How does the system "do" an action? In the part of the foreground periphery where input arrives from the motivational/drive system, there is a unique spot—the "make-it-so" place—that drives all actions. If an action atom can get itself attached here, this triggers an outward-moving relaxation wave that will (usually) result in signals being sent from the motor output area that cause muscles to do something. The make-it-so place, in the part of the foreground periphery we have called the motivation area, is the place where the buck starts.

Neural Implementation

Nothing has been said, so far, about how this framework is realized in the brain's neural hardware. There are many possibilities here, because this framework is primarily designed to specify high-level mechanisms. The framework itself is neutral with respect to neural implementation.

With that qualification, what follows is a quick sketch of one way it could be realized in the brain. This neural implementation will be assumed in the rest of the chapter.

The cortex could be an overlapping patchwork of "processors," each of which can host one active atom, and the sum total of all these processors is the thing that we have called the foreground. Each processor is a large structure with quite complex functionality, and is capable of doing such things as hosting a particular atom for a while, transferring a hosted atom to an adjacent processor if there is pressure for space, and setting up rapidly changing communication links to other atoms located some distance away across the cortex. One possibility is that the central core of each processor corresponds to a cortical column.

The generic, passive atoms that are stored in the background (long-term memory) are colocated with the processors that were just described, but each processor can hold a large number of generics. Each of these passive atoms is encoded in distributed form inside the processor. At the level of the entire cortex, then, a generic atom would seem localized (because it is within one processor), but since each processor is quite extensive, the atom is not at all localized within the processor.

The activation of an instance atom involves a call to the processor that hosts the generic, which causes the processor to find a spare processor that can host the instance atom. The parent processor itself might be able to play host, but if not, then it passes the atom to some other processor (possibly a neighbor) for hosting. One way or another, an activated atom is quickly copied into a processor that can handle it, in much the same way that a computer program might be transferred across a network to a computer that can run it.

In summary, then, the cortex can be viewed in two completely different ways. It can be seen as the foreground, in which case it is effectively a space within which the extremely volatile

instance atoms arrive, set up a rapidly changing set of links to other instance atoms, and then depart. Alternatively the cortex can be seen as the place where the contents of long-term memory are stored, since the generic atoms are also located in the processors.

Example 1: Visual Object Recognition

We conclude this summary of the molecular framework with two examples of the kinds of activity that go on in the foreground. The first involves the recognition of an object perceived with the eyes, and the other is the carrying out of an action.

Suppose the system were to start in a thought-free, meditative state, in a darkened room, and then suddenly a light comes on and illuminates a single chair. The foreground would start out relatively empty, and when the light is turned on a sequence of atoms would come into the foreground until, at the end of the process, the [chair] atom would be activated.

This process would begin with the arrival at the foreground periphery of signals along the lines coming from the visual system. When these signals arrive, they trigger the activation of instance atoms representing some low-level features of the visual input. When these atoms appear in the foreground, they attach to the places on the periphery where their features occur, and then they look around at their closest neighbors. If a pair (or perhaps a group) of these atoms recognize that they have occurred together before, they will know about some other atoms that they could call into the foreground, which on those previous occasions represented their co-occurrence. They will activate those second-level atoms, and these in turn will try to attach themselves strongly to the first-level atoms. The ones that succeed in making bonds will then feel confident enough to look around at their neighbors and repeat the process by calling in some even higher-level atoms. An inverted tree of atoms thus develops over the part of the foreground periphery where the chair image came in, and at the top end of this inverted tree, finally, will be the [chair] element.

This picture of the recognition of a chair is extremely simplified, but it gives a rough picture of the kind of activity that occurs: atoms being activated by others, then each atom trying to fit into a growing structure in a way that satisfies its own internal constraints about the roles it can play. As a whole, the system tries to relax into a state in which some atoms have self-consistently interpreted the new information that arrived.

Example 2: Sitting Down

Suppose the person who just recognized the chair next hears the experimenter ask them to sit down. The atoms representing this request will (after an auditory recognition process very similar to the visual recognition event previously described) bump into, and interact with, the large cluster of atoms hovering around the motivation area of the foreground periphery. This cluster represents the mind's complex stack of goals and intentions, all the way from its most nebulous motivations (stay safe, seek warmth, get food, etc.) to its most specific action schemas (accept this experimenter as a trusted friend whose requests should be obeyed). As a result of this interaction between the atoms representing the request and the atoms around the motivation area, a new atom that encodes the sitting action is activated and attached to the make-it-so place on the foreground periphery, and this gives the [perform-a-sitting-action] atom enough strength to cause a cascade of other atoms to be brought in, which then elaborate the sitting plan in the context of where the body currently is, where the chair is, and so on.

In the case of the sitting-down action, a wave of relaxation emerges from the [perform-a-sitting-action] atom and causes a sequence of atoms to spread toward the motor output area. When these atoms arrive at the periphery, muscles move and the sitting action occurs. The only difference between this and visual recognition is that the wave of relaxation does not come from the sensory input area and end with the activation of the [chair] atom, but starts with one atom at the make-it-so place and ends with a broad front of new atoms hitting the motor output area.

CONCLUSION

If complex systems are what they seem to be, then the universe contains some systems that are impenetrable to scientific analysis, in the sense that we can observe their behavior but cannot develop any kind of analytic proof that this behavior is the result of the underlying mechanisms. If this impenetrability sometimes occurs in systems that are partially complex, but also partially non-complex, we could find ourselves in a situation where we first explain some aspects of that system, but then convince ourselves that the rest of the system will eventually fall to the same scientific attack. Unfortunately, the stubbornly complex aspects of the system could resist attack for a very long time, because the mechanism behind those aspects might look utterly unreasonable—it might be the kind of mechanism we would never have guessed would be responsible.

When the ramifications of this idea are examined in depth, it appears that our approach to cognitive science—the entire methodology we use for unlocking the secrets of human cognition—might be in need of drastic revision.

The revision proposed in this paper is to find ways to build large sets of explanatory models, rather than just single models, and to insert these into simulations that can then be used to explore how all of these candidate models behave. In this way, we open the door to considering models that look unreasonable on the surface, but which may in fact be the only viable explanation for a given set of experimental data.

Postscript: The Need for New Software Tools

If this new approach to cognitive science is to be implemented, one of the first prerequisites will be software tools capable of building models and organizing them into simulations. In the past, cognitive researchers have tended to build small pieces of software to implement their models, and this process has required them to be part-time software engineers as well as psychologists. The results have been mixed: such models are often very simple, and incapable of generalization. It would be impossible to expect the proposed new approach to cognition to be implemented unless researchers could be liberated from the burden of tool low-level programming.

Historically, the arrival of new tools has often been the vital catalyst that starts technological revolutions. A lack of the right tools can perhaps be seen as the single biggest factor that has caused the complex systems problem to go unrecognized for so long: with no way to do anything about it, there is little incentive to consider it. What is needed now, then, is the kind of software that might trigger a new cognitive revolution.

REFERENCES

Ackley, D.H., Hinton, G.E. and Sejnowski, T.J. (1985) "A learning algorithm for Boltzmann machines," Cognitive Science 9:147-169.

David E. Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) "Learning representations by back-propagating errors," Nature 323:533-536.

Dux, P. E., Ivanoff, J. G., Asplund, C. L., & Marois, R. (2006). Isolation of a central bottleneck of information processing with time-resolved fMRI. Neuron, 52, 1109-1120.

Gardner, M. (1970) "Mathematical Games: The fantastic combinations of John Conway's new solitaire game 'life'." Scientific American 223(4): 120-123.

Gorman R.P. and Sejnowski, T.J. (1988) "Analysis of hidden units in a layered network trained to classify sonar targets," Neural Networks, 1(1), 75–89.

Guy, R. K. (1985) "John Horton Conway," in Albers and G L Alexanderson (eds.), "Mathematical people: Profiles and interviews." Cambridge, MA: 43-50.

Harley, T. A. (2004). "Does cognitive neuropsychology have a future?" Cognitive Neuropsychology, 21, 3-16.

Harley, T. A. (2004). "Promises, promises. Reply to commentators." Cognitive Neuropsychology, 21,51-56.

Holland, J. H. (1998) "Emergence." Helix Books, Reading, MA.

Horgan, J. (1995) "From complexity to perplexity." Scientific American 272(6): 104-109.

Kohonen, T. (1987) "Self-organization and associative memory." Springer: Berlin.

Kuhn, T.S. (1962) "The structure of scientific revolutions." University of Chicago Press, Chicago, IL.

Loosemore, R.P.W. & Harley, T.A. (2010). Brains and Minds: On the Usefulness of Localisation Data to Cognitive Psychology. In M. Bunzl & S.J. Hanson (Eds.), Foundational Issues of Neuroimaging. Cambridge, MA: MIT Press.

McClelland, J.L., D.E. Rumelhart and the PDP Research Group (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models, Cambridge, MA: MIT Press

McClelland, J.L., Rumelhart, D.E. & Hinton, G.E. (1986) "The appeal of parallel distributed processing." In D.E. Rumelhart, J.L. McClelland & G.E. Hinton and the PDP Research Group, "Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1." MIT Press: Cambridge, MA

Mitchell, M. (2008). Complexity: A Guided Tour. New York: Oxford University Press.

Rumelhart, D.E., J.L. McClelland and the PDP Research Group (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations, Cambridge, MA: MIT Press

Russell, S. J. and Norvig, P. (1995) “Artificial Intelligence: A modern approach.” Prentice Hall, Upper Saddle River, NJ.

Waldrop, M. M. (1992) “Complexity: The emerging science at the edge of order and chaos.” Simon & Schuster, New York, NY.

Wolfram, S. (2002) “A New Kind of Science.” Wolfram Media: Champaign, IL. 737-750.