The Complex Cognitive Systems Manifesto

Richard P. W. Loosemore
Mathematical and Physical Sciences, Wells College, Aurora, NY 13026, U.S.A.
rloosemore@wells.edu

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 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. And this need for clarity in the mapping from high level descriptions to low level mechanisms is not unique to brain science: there are many systems in the world—both natural and artificial—that are so complex that different types of explanation exist at different levels, and in each case it is vital to understand how the levels relate.

This is especially true in the context of nanotechnology and the brain. We are approaching a time when nanotech tools will allow us to intervene, augment or duplicate brain systems: but the mere possession of a microscopic screwdriver is not the same thing as knowing what will happen when said screwdriver is inserted into the circuit board of a supercomputer. Wielding the screwdriver must be backed up with theoretical knowledge of how the system works.

The poster child for high-level to low-level mappings must be the science of chemistry. We have a more-or-less coherent picture of chemical processes that starts down in the quantum physics of atoms and goes all the way up to the functional properties of enzymes. But this story about different levels of description in chemistry was assembled by an extremely reductive program of research in the hard sciences, and in this paper I am going to suggest that there is a serious structural problem with using that same type of reductionist scientific methodology to understand the relationship between high and low levels in the human cognitive system. In spite of the seemingly rapid advances being made in neuroscience and

cognitive psychology—and especially in the field of brain imaging—I would argue that progress is far less significant than it might appear because there is 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. Moreover, it seems 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.

To get a bird's-eye view of what the complex systems problem involves, imagine a dynamic system that contains many interacting components, and imagine that the net result of the component interactions is that the overall behavior of the system shows some noticeable regularity. Now suppose that there *does not exist* any mathematical analysis that will allow someone to start from the component interactions and predict the appearance of that overall regularity. Under such circumstances a number of serious problems would arise if researchers were to treat the system in the same way that "regular" systems (where there is almost always an underlying mathematical model waiting to be discovered) are treated. That set of serious problems is what is meant by the *complex systems problem*.

Rather than try to define the complex systems problem in its most general and abstract form, this paper will focus on how the CSP impacts one particular aspect of cognition: how behavior at the cognitive level should be explained in terms of events at the neural level. In other words, the neural-cognitive mapping.

The neural-cognitive mapping is the backbone of all cognitive science. The mapping question 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 experiments can hardly be interpreted at all without making assumptions about how the psychological level is related to neural events.

There are two main reasons why it is important to present 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—variously known as *connectionism*, *neural nets* or *parallel distributed processing*—which turns out to have been a textbook illustration of how the complex systems problem can undermine research. 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 compromised 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 is perpetuating 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. What defines a regular systems is the fact that we can write down equations or algorithms that describe how the components of the system interact, and these equations or algorithms can be solved to get a prediction of the system's overall behavior. As it happens, most of the systems of interest to science belong in the regular 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 its 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, a complex system has an overall behavior that can only be understood by constructing a simulation of the system. If we simulate the system's underlying mechanism and then observe that the simulation behavior corresponds to the original system behavior, we might say 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 that the overall behavior must arise, given those underlying mechanisms. Unfortunately, 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.

This is an extremely simplified (even 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 into the detailed analysis, it might be worth summarizing the shape of the full argument that will eventually emerge:

- 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 as if they could never explain the behavior.
- Because of this difficulty, the process of finding a scientific explanation for a partially complex system will exhibit a kind of *global pathology*—which is to say, if our scientific methodology is driven by a search for locally-plausible models for every aspect of the system, we will never be able to integrate these locally-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-plausible models: but if cognitive systems are partially complex systems this strategy will result in an endless stream of local models that never fit together.

The main conclusion to be had 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 a group of people trying to lay square tiles on a floor that appears flat if examined locally, 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, it might seem that these gaps can be resolved by adjusting any pair of conflicting tile patches. In normal space these pairwise adjustments would eventually converge on a global solution, but if the space is curvilinear, these pairwise changes will never converge on a solution.

Characterizing Complexity

This might seem to be a good point at which to give the accepted definition for the term "complex system." Unfortunately this is 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 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 what effect 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 the thing that is most apparent, 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 proper, and the *form* (shape, structure, or state) of the system. The first is a dynamic feature, the second is less time-dependent. For the sake of narrative convenience

the term *behavior* will often be used to signify both of these. So "behavior" can be used to cover both the dynamic and static features of the system.

Regularities

When we talk about a system's behavior what we usually 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 like the role played by hurricanes in the world's ecosystem, or the typical regions in which they occur, or the typical time history of a 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. So for example, we would say 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 cannot always be 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 writing down 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. This means that 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 very significant but 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 an example 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, divine intervention or blind guesswork, but by being sensitive to many factors that pointed toward the correct mechanism.

Forward Path as One-Way Street

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 only seems 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 a sign that somewhere beneath the surface there exists an elegant explanation for the behavior. But however compelling this might seem—and however frequently it turned out to be true 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 troublesome question, because we can never be sure that *no* explanation will ever be possible. We might suspect a system of being beyond explanation in this way, but there is always a chance that we will be surprised, tomorrow, by an unexpectedly new and elegant proof.

What we can say, however, is that 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. Another way to phrase this is that for all practical purposes we have to assume that no rigorous, analytical explanation will be possible.

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 in the usual way, 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 will 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 is 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 that we study artificial 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 and 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" that explains the behavior), 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 (the process of intuiting a 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 will not come into play. 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 might be possible for a simulation to 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 given a great deal of attention because, as I explained above, nobody tries to invent complex systems that have a pre-ordained behavior. That kind of choose-the-behavior-first 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 the behavior (effect), the mechanism (cause), or the connection between the two. Notice in particular that a system is never complex *because* it has certain behaviors, or certain mechanisms—rather, it is the relationship between these two that makes it complex. This is a frequent point of confusion in discussions of complexity. Calling a system "complex" is really a statement about whether we have any chance of discovering a theory that concisely describes what the behavior should arise from the mechanism.

Having said that, though, if we look at the empirically observed characteristics of known complex systems, it is possible to give a list of design ingredients, or aspects of the mechanism, that tend to make a system complex. If we see a system in which a plurality of these features occur, we could say that past experience teaches us that a complex relationship might exist between behavior and mechanism:

- The system contains large numbers of interacting 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.

No Diagnostic Test for Complexity

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. But although conceivable, 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 complex-systems 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 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 complexity effects ever became 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. Denial is easy in this case 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 already seem to have 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 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 appear 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 can be built inside the system.

Other ways to encounter partial complexity are easier to appreciate. A system can have a number of subsystems, 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.

Complexity exists as a continuum, so we could say that some partially complex systems are *primarily regular*, while others are *dominated by complexity*. The solar system would be primarily regular, because there are simple 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 (prototypes?, exemplars, clusters 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 at 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 the purpose of this line of argument. 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 be willing to take the next step and admit that this has great significance, but it is enough for our purposes to note that the existence of complexity in this context is widely accepted.

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, such an argument would be supported by a proof that cognitive systems must be complex systems, so 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 in the known properties of the concept mechanism is sufficiently high that it is imperative that we ask whether the consequences of that complexity would be 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 choose to investigate 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, one practical consequence 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 and the backward directions. If the link is broken, the mechanism must look unreasonable in some way.

But this means 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—models in which there is always an understandable relationship between the behavior being explained and the model being used to explain it—then we will be setting ourselves up for failure. The system cannot be entirely made from components that have a non-complex relationship to the behaviors they produce. If the system is partially complex, it must be the case that at least some of the models have a pathological (complex) relationship to the behaviors they generate. We may never know which components should have this pathology (although we can sometimes make a shrewd guess), but we can be sure that if we have prior reasons to suppose that the system 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 combine 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 doomed strategy: it will always miss the truth.

What would happen if, in spite of the danger, we simply forged ahead and applied the usual scientific strategy? The naive conclusion might be that in the case of those system components that require a complex explanation, we would see our models breaking and eventually conclude that this component was the one that needed to be treated differently. Alas, there is no reason why the locus of trouble should be so easy to pin down. 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 would probably 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 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 successful 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 efforts at 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, the goal would be unattainable, and the separate model-building efforts would only become more complicated and less plausible as time progressed.

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

Solving the Problem

The main way to start solving this problem is to avoid a narrow focus on locally optimal models. To do this we need to find ways to develop the widest possible range of different models for each component of the target system, where previously we might have chosen only one. Then these models need to be tested in parallel, because it is important to assume from the outset that the proper explanation for any given behavior could be a mechanism that looks unreasonable. To some extent this involves a semi-blind search through the space of all possible explanatory models.

Inventing single models is hard enough, but 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, but automation implies that we need to stop thinking of models as individual hand-crafted works of art, and instead treat them as entities that can be described in terms of *design parameters*. Then, with this new vision of a model as a commodity—as just a set of choices for the design parameters—we can start to produce large numbers of models, each of them being a single point in a multidimensional space of parameters.

This would entail a radical shift of mindset, from models to *generators* of models. Instead of thinking of models as separately interesting things, we need to think about the kind of engines that could be used to generate sets of models.

This would clearly bring a significant change to cognitive science. There would still be room for cognitive scientists to interpret the results of a particular 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 to see if new parameter-concepts need to be invented to capture it.

Frameworks and Paradigms

One possible approach would be to write an abstract mathematical formalization of the space of possible cognitive systems, and then devise an algorithm to explore that space. This might

be called the "pure mathematics" approach to the problem. Although it might be attractive to mathematicians, such an approach seems less than optimal. After all, we do not expect every aspect of the target system to be egregiously complex, so there will be many features of the its behavior that we can deduce from regular scientific analysis. Going back to first principles and writing abstract formalisms to describe everything in the system would be equivalent to throwing away our existing knowledge.

The main alternative to the pure mathematics strategy would involve somehow collecting all of the existing knowledge about the human cognitive system into one large framework, and then using that framework as the basis for a parameterized exploration of different models and components that are consistent with the framework.

The term *framework* is used here in the particular sense that is common in the philosophy and methodology of science: a framework is a loose set of organizing ideas and assumptions that inform a cluster of theories. Individual theories derived within a framework are supposed to be subject to confirmation or refutation as the result of experimental data, but the framework itself exists at a higher level, and is not subject to direct empirical attack. The concept of a *paradigm* (Kuhn, 1962) is closely related, although it could be argued that frameworks are situated at a level of generality somewhere in between paradigms and theories.

The traditional separation of power between models/theories, on the one hand, and frameworks on the other, is that the models/theories are dominant while the frameworks sit quietly in the background. Papers get published because they supply new data that is pertinent to some models or theories that are supposed to account for the data; papers rarely get published if they claim to deliver an improvement at the framework level. Success is immediately verifiable in the former case, whereas claimed improvements in frameworks are often deemed to be mere speculation.

In order to solve the complex systems problem, this situation has to be inverted. It makes no sense to *always* put a premium on models that are good at explaining particular data, or on research that discriminates between pairs of such models. Instead, what matters is our ability to generate models as commodities within a framework. So, rather than leave the framework in the background as a poorly articulated cluster of ideas, the framework needs to be brought front and center, where it can be made as explicit as possible and turned into a mechanism that generates (hopefully, in an automatic way) a very large set of candidate models that can then be implemented and tested for their fit to the data.

The problem, of course, is that saying we need fully articulated frameworks is not the same thing as actually building them. For example, what is the framework or paradigm currently accepted by cognitive psychologists as their consensus assumption about the way cognition happens? A cursory glance at any standard textbook of cognitive psychology makes it clear that the field consists of many different sets of ideas, some of which may be backed by profoundly incompatible frameworks. What, for example, are the common assumptions about cognitive processing shared by the cohort model of spoken word recognition (Marslen-Wilson, 1990) and the Bruce and Young (1986) model of face recognition? If the McClelland and Rumelhart (1981) model of word recognition is considered valuable because of its sub-symbolic aspects, how does it square with the apparently symbolic process that occurs in sentence production?

Frameworks as Art

There is an uncomfortable truth at the heart of the framework-building process. Frameworks are not constructed by cautious logical arguments and scrupulous support from empirical data. They are intuited. They represent extended judgment calls. They are born out of a collective feeling of what is elegant and right, and what most comfortably sits with the various local ideas that are working best at the moment. Sometimes a framework has to be a bold leap in the dark.

This is another way of saying that if we accept that the complex systems problem must be solved by a new emphasis on top-down design—starting with a big-picture view of the human cognitive system, turning that into a concrete framework, then exploring very large numbers of models within the framework—we must also accept that there can be no hard-and-fast rules for how to devise the framework. There is a need for a new kind of theoretical activity within cognitive science, in which frameworks are invented and described in the broadest possible terms, and where the scientific community judges them on the degree to which they appeal to a sense of what is elegant.

The most viable frameworks should then be turned into explicit software engines that enable models to be constructed and tested. But even though the process should eventually lead to proper empirical testing of the sort that science is familiar with, the initial process of framework construction needs to be given the breathing room it needs to be creative and inventive. There needs to be an attitude shift, so that a clear line exists between allowed creativity and invention on the part of the framework theorists, and down-to-earth Darwinian ruthlessness of the sort that applies when specific models are being confirmed or refuted by empirical data. The framework theorists need credibility and respect.

Connectionism and Constraints

I will close by showing how an explicit framework can emerge from the weaknesses of an older, less explicit and non-CSP-friendly framework.

The older framework was known as *connectionism* or *parallel distributed processing* (McClelland, Rumelhart & the PDP Research Group, 1986; Rumelhart, McClelland and the PDP Research Group, 1986). The connectionist revolution that happened in the 1980s was often taken to be about neuron-like processing units, but although this was an important feature of the new ideas, the background motivation was actually more general than that. It was about finding ways to build cognitive models in which *multiple simultaneous constraint relaxation* was the dominant theme. McClelland, Rumelhart 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 aspects of cognition is that they seem to point toward a type of model that involves large numbers of objects connected by weak constraints, in such a way

that none of the constraints is rigidly enforced, but where the system nonetheless shows intelligent behavior. The challenge faced by the early connectionists was to find ways to build such models, and what was inspiring about the connectionist revolution was that a group of early algorithms—Boltzmann machines, interactive activation, back-propagation, among others—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 their global behavior. So the primary characteristics of these systems were not complex.

But if the CSP is real, 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 large 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. The field tended toward stagnation, with no dramatic solutions to the larger problems of cognition to follow the bold progress that occurred within a few years at the start of the revolution.

The field was driven by a strong, implicit assumption that the best (and perhaps only) place to look for models in which constraint satisfaction was the driving force, was in 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

So if we wanted to hit the restart button on the connectionist revolution, exploring models of cognition in which mutual simultaneous constraint relaxation played a significant role, what kind of framework might we use? There are many possibilities, of course, but in what follows I will articulate 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 we might go next, now that we understand the force of the complex systems problem.

This *molecular framework* (Loosemore and Harley, 2010) has one simple idea at its core. When the relative strengths and weaknesses of connectionism were first debated, many of the weaknesses could be traced to the fact that the neuron-like units doing the constraint relaxation were locked into fixed positions within the system (just as real neurons are). This eventually leads to trouble, because we know that higher cognition involves clusters of concepts that can be rapidly linked together in different configurations, and fixed neurons cannot easily do this unless every one is connected to all others. These configurations of concepts are indispensible: it is not enough to know that the nodes for [girl], [bites] and [dog] are all active: what matters is whether they are arranged in the sentence "girl bites dog", or "dog bites girl". Cognition also involves situations in which there are copies of the same concept: the sentence "dog eat dog" needs to be represented with two instances of the

dog concept. So, in order to model these processes, we would like to have a system in which multiple copies of concepts can be activated, with flexible, transient relationships between those concepts.

This seems to imply a return to the older, pre-connectionist ideas that involved symbol processing—where the symbols are free to be created, deleted, linked and modified in any way, and perhaps with structure inside the symbols—but with the additional feature that the symbols have the type of mutual, simultaneous constraint relaxation properties that are the hallmark of a connectionist system.

The fundamental construct of the molecular framework, then, is a symbol-like entity that is free to roam around like an atom in a molecular soup. As the atom moves 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 leads to a picture in which transient, ephemeral molecules are being made and unmade—each molecule resembling a structured model of some aspect of the world.

The goal now is to see whether a cognitive framework can be built around this generalized version of connectionism, in such a way that most aspects of human cognition make sense in the framework. The following is a brief sketch of some of the main features of the proposed framework.

- Symbols appear in two distinct locations: they are either in long term memory, where they do nothing, or a copy of the symbol is transferred to a special place (the *foreground*) in which the system is representing the current state of the world, and all of its currently active thoughts. To make it clear that these really are distinct, the symbol is referred to as an *element* when it is in long term memory, and as an *atom* when it is in the foreground. There is only one element for each symbol, but there can be many atoms.
- To make this distinction more concrete, we can make a tentative mapping between these abstract notions and the neural hardware of the brain. Suppose that the primary functional role of a cortical column is to host a single active atom. A secondary role of the columns is to keep a collection of elements, in such a way that each column keeps an overlapping subset of the total set of elements in the system's long term memory. A single element might be redundantly encoded in a cluster of (perhaps adjacent) columns, so it could not be destroyed by a malfunction in any one of the columns. When a particular element is called upon to deliver a copy of itself—in the form of an atom—one of these host columns produces the atom and tries to find a place for it on the nearest column that is either vacant, or which contains the weakest active atom. Thus, as time goes on the population of atoms hosted across the entire sheet of cortical columns will change in such a way that the strongest atoms stay active and the weakest are replaced by newly activated atoms. When an atom is deactivated it is returned to the original element, where some aspects of its recent activation are used to update the element.
- The concept of working memory is therefore roughly equivalent to the current set of atoms being hosted on all cortical columns (the total number of columns being of the order of one million—although this depends on how columns are counted).
- When an atom is active its job is to look at the relatively nearby atoms and try to find ways to get support by establishing connections to them. To this end, every symbol has

an internal idea of what kind of neighborhood it would like to see in the foreground. The neighborhood is not necessarily a simple list or fixed set of other atoms, but can contain slots (variables) that can be filled by atoms that themselves must satisfy certain requirements. If an atom can quickly establish links to neighbors in such a way that both it and they find the link satisfies their internal desired-neighborhood map, the atom becomes strongly supported and is less likely to be deactivated.

- One of the main sources of support for foreground atoms is the flood of information coming in from the senses. This information drives certain columns to activate certain atoms that represent the incoming information. So there would be some columns whose hosted atom is determined, not by the presence of nearby supportive atoms, but by the fact that it has a direct line to (say) a patch of color detectors in the retina.
- Activation of high-level representations would then proceed in a manner similar to that which occurs when an input pattern arrives at a neural network: neighboring, strongly supported atoms that encode low-level sensory signals get together to active candidate atoms that might represent slightly higher symbols. This bottom-up cascade of activation happens in the context of some top-down activation of atoms representing expected features.
- Notice, at this point, that the framework does *not* specify exactly how the mutual interactions between atoms will cause clusters to form, representing coherent thoughts or coherent representations of perceived objects: the exact choice of mechanisms is expected to be especially sensitive to complex system effects, so this is the area in which exploration of different possible mechanisms has to be undertaken in an empirical manner. To cover our future discoveries of the exact mechanisms, we simply summarize this aspect of the system by saying that what the foreground atoms do is engage in mutual, simultaneous, weak constraint relaxation, in such a way that the system as a whole exhibits coherent, intelligent thought processes.
- The interactions between atoms may include a number of unorthodox features (at least, unorthodox in purely connectionist terms). Some atoms may act as operators, whose only function is to transform configurations of other atoms in certain ways, then leave (there may be loosely defined set of *analogy* operators that do this on a routine basis). Also, a large number of atoms will represent *relationships* rather than features, with hierarchies of these relationship atoms that are every bit as extensive as the hierarchies of features. And there may be some processes that are not encapsulated in atoms at all: for example, a process of elaboration that sometimes kicks in when a cluster of atoms fail to reach consistency with respect to one another: the initial failure causes the elaboration level to go up, so that all participating atoms try to surround themselves by greater-than-normal numbers of related atoms.
- The concept of an elaboration area can be used to capture the idea of the *focus of attention*. At any given time there will be a unique zone of high elaboration that is the attentional focus, with special mechanisms that cause it to move from place to place in the foreground, depending on factors like the unexpectedness or novelty of certain sensory inputs, or sequences of learned actions.
- Just as there is a flow of activation from the arriving sensory input, as more and more abstract symbols are activated to represent the input, so there is a dual version of this process going on toward the motor output end of the foreground. The flow starts in an area of the foreground that might be labeled the Make-It-So region, which is the primary

source of support for the tree of atoms that eventually lead down to the specific muscle signals required to execute a movement.

Molecular Framework: Summary

This short account of the molecular framework is intended only as a hint of what might be done to start a research program that is immune to the complex systems problem. The next stage would be to elaborate this account in enough detail to see how, in principle, various specific cognitive phenomena could be accounted for, then implement this as a software (and perhaps hardware) system designed to support a class of models consistent with the framework. After the software is built, large numbers of different candidate mechanisms need to be set up within the system and their properties explored under a variety of circumstances. An iterative process of exploration and analysis would then (hopefully) lead to a convergence on systems that are good approximations to the particular system that is used by the human brain. At that point, we might be able to say that we understand the neural-cognitive mapping in enough detail to know what kinds of nanotech interventions would have which effects on the system.

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 Urgency of 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 (even simplistic), 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 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 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 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.

Bruce, V., and Young, A.W. (1986). Understanding face recognition. British Journal of Psychology, 77, 305-327.

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

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

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

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.

Marslen-Wilson, W.D. (1990). Activation, competition, and frequency in lexical access., In G.T.M. Altmann (Ed.), Cognitive Models of Speech Processing: Psycholinguistics and Computational Perspectives (pp.148-172). Cambridge, MA: MIT Press.

McClelland, J.L. & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. Psychological Review, 88, 375–407.

McClelland, J.L., D.E. Rumelhart & 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 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.

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.