Complex Systems, Artificial Intelligence and Theoretical Psychology

Richard LOOSEMORE
Surfing Samurai Robots, Genoa NY, USA
rloosemore@surfingsamurairobots.com

Abstract. The main finding of complex systems research is that there can be a disconnect between the local behavior of the interacting elements of a complex system and regularities that are observed in the global behavior of that system, making it virtually impossible to derive the global behavior from the local rules. It is arguable that intelligent systems must involve some amount of complexity, and so the global behavior of AI systems would therefore not be expected to have an analytic relation to their constituent mechanisms. This has serious implications for the methodology of AI. This paper suggests that AI researchers move toward a more empirical research paradigm, referred to as "theoretical psychology," in which systematic experimentation is used to discover how the putative local mechanisms of intelligence relate to their global performance. There are reasons to expect that this new approach may allow AI to escape from a trap that has dogged it for much of its history: on the few previous occasions that something similar has been tried, the results were both impressive and quick to arrive.

Keywords. Complex Systems, Artificial Intelligence, Adaptation, Self-Organizing Systems, Cognition, Methodology, Theoretical Psychology, Cognitive Science.

Introduction

One the most basic assumptions made by Artificial Intelligence researchers is that the overall behavior of an AI system is related in a lawful, comprehensible way to the low-level mechanisms that drive the system.

I am going to argue that this apparently innocent assumption is broken, because all intelligent systems, regardless of how they are designed, must be complex systems[1], and the definition of a complex system is that the overall behavior of the system is not always related in a comprehensible way to the low level mechanisms that cause the behavior. I will further argue that the consequences of this assumption being broken are so serious that the current approaches to AI will never lead to a full, human-level artificial intelligence. The good news is that there is a way to solve the problem, and this solution could unlock the floodgates of progress in AI. The bad news is that many AI researchers are so uncomfortable with this solution that they are determined to resist it at all costs. My goal in this paper is to discuss the problem, outline the solution, and try to raise the awkward issues surrounding the negative reaction that the solution tends to provoke.

This is the "Complex Systems Problem," and in the second part of the paper I propose a new methodology designed to overcome it. The new methodology involves a

[1] Unless otherwise noted, the terms "complex" and "complexity" will always be used here to refer to aspects of complex systems, not to the general sense that means "complicated," nor to the mathematical analysis of algorithmic complexity.

combination of three factors: (i) a grassroots integration of AI and the field of cognitive psychology, (ii) a "complete framework first" approach to AI system design, and (iii) the use of programs of systematic experimentation, in which large numbers of systems are built in order to discover (rather than assume) the relationships between global behavior and local mechanisms.

It is important to emphasize that the proposed solution is rather more than just a new approach to AI. It is based on a body of knowledge (about human cognition) and a philosophical attitude (about the unavoidable need for empirical science) that is categorically rejected by many of the researchers who now dominate the AI field. The awkward truth is that the proposed new methodology has little use for most of the work that has been done in the AI field in the last two decades, and even less use for the devotion to formal systems that seems indispensible to many in the AI community.

These considerations about the social psychology of research in AI are important because they underline the extraordinary seriousness of the Complex Systems Problem. If the only way to solve the problem is to declare the personal philosophy and expertise of many AI researchers to be irrelevant at best, and obstructive at worst, the problem is unlikely to even be acknowledged by the community, let alone addressed.

The ultimate conclusion of this paper is that the Complex Systems Problem is the single biggest reason why a human-level artificial intelligence has not been built in the last fifty years, but that if someone has the vision and resolve to fly in the face of orthodoxy and implement the solution proposed here, the results could be dramatic. Quite apart from the fact that this approach has never been tried before, there are positive reasons to believe that it would succeed where conventional AI has not.

1. Complex Systems

This section and the next examine the nature of complex systems and their impact on AI research.

A complex system is one in which the local interactions between the components of the system lead to regularities in the overall, global behavior of the system that appear to be impossible to derive in a rigorous, analytic way from knowledge of the local interactions.

This definition is part empirical fact, part mathematical intuition. As empirical fact, it is a summary of observations made independently by a number of researchers who tried to understand the behavior of systems in a wide variety of fields [1].

The mathematical intuition aspect is perhaps more powerful, because it derives from something that has been said by nonlinear mathematicians for a long time: taken as a whole, the space of all possible mathematical systems contains a vast, unexplored region in which the systems are easy enough to define, but seem to be beyond the reach of analytic solution. It would probably be no exaggeration to say that the part of this space that contains the sum total of all systems that have so far been rigorously analyzed is considered by many to be an unthinkably small corner of the whole space. The crucial part of the intuition concerns how far the tractable corner of this space might expand in the future: if human mathematicians could work for an arbitrary length of time, would they be able to find analytic solutions for most of the systems in the unexplored region? There is no particular reason to assume that we could do this, and it would not be surprising if there exist no analytic solutions whatsoever that describe the behavior of almost all of the systems in that space.

The implication of this intuition is that if a system has components that interact in such a nonlinear, tangled way that we have doubts about whether any analytic understanding of its behavior will ever be possible, we should at least adopt the precautionary principle that there might not be any analytic explanation of the system. We should be especially careful when we observe regularities in the global behavior of such systems: those regularities should not be taken as a clue that a formal, analytic explanation is lurking beneath the surface. The term "complex system" is used to describe precisely those cases where the global behavior of the system shows interesting regularities, and is not completely random, but where the nature of the interactions between the components is such that we would normally expect the consequences of those interactions to be beyond the reach of analytic solutions.

This defining feature of complexity is sometimes referred to as "emergence" [2], but since that term has wider significance (and is even taken by some to have vitalist connotations that border on the metaphysical), the term "Global-Local Disconnect" (or GLD) will be used here. The GLD merely signifies that it might be difficult or impossible to derive analytic explanations of global regularities that we observe in the system, given only a knowledge of the local rules that drive the system.

1.1.The Recipe for Complexity

Speaking in purely empirical terms, 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 system activity does not go into a locked-up state where nothing changes, and the activity does not go into an infinite loop, then there is a high probability that the system will show signs of complexity.

It should be understood, however, that this recipe only gives a flavor of what it takes to be complex: it is possible to find systems that lack some of these features but which would still be considered complex.

1.2.The Absence of a Clear Diagnostic

One fact about complex systems is especially important in the context of the present paper, where a great deal hinges on whether or not AI systems can be proven to have a significant amount of complexity in them:

- 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 global regularities that, if the system is complex, cannot be derived analytically from the local features of the system.

What this means is that any debate about whether or not intelligent systems are complex must not involve a demand for a definitive proof or test of complexity, because there is no such thing.

The "virtually" qualifier, above, refers to the fact that complex systems are not excluded from having global properties that are derivable from local features or mechanisms: it is just that such exceptions to the GLD are unusual, especially where, as in this case, the property is supposed to hold across a large, nonhomogeneous class of systems.

1.3.Controversy

The scientific status of complex systems research has been disputed. Skeptics like John Horgan [3] have argued that the term "complex system" means whatever people want it to mean, and that in spite of all the effort put into the field there is little in the way of a theory that might unify the disparate phenomena under study.

This criticism may have some merit, but it is important to understand that Horgan's points have no relevance to the arguments presented here. It is not important that there be a general theory of complexity, because all that matters here is one rather simple observation about the behavior of complex systems: the global-local disconnect. The GLD is a deceptively innocent idea—it seems to lack the kind of weight needed to disrupt an entire research program—but if it is real, and if complexity plays a significant role in the behavior of intelligent systems, then much of mainstream AI research may nevertheless have been severely compromised by it.

1.4.Computational Irreducibility

Could it be that complex systems only appear to have a global-local disconnect because current mathematical tools are not up to the task of analyzing them? Perhaps in the fullness of time these systems will be understood in such a way that the global regularities are derivable from the local mechanisms? In that case, the GLD might be nothing more than temporary pessimism.

This flies in the face of the intent of the complex systems idea, which is that something more than mere pessimism is involved; there really is something new and fundamentally different about many of these systems, and there will not be any future mathematical formalism that eliminates the GLD.

Stephen Wolfram has used the term "computational irreducibility" [4] to capture one version of this idea. Physical systems obey laws that govern their behavior, and Wolfram's suggestion is that those laws can be viewed as if they were computations being carried out by a physical system in order to calculate the next set of values for its physical state. What we have known ever since the time of Isaac Newton is that we can reduce these laws to equations that allow us to short-circuit the physical computations and predict what a physical system will do before it does it. If we want to know where a planet will be next week, we can use equations that do not take a whole week to complete, whereas the computations that the planet is effectively executing when it follows its local laws of nature do take a week.

What Wolfram points out is that although we have had great success in finding mathematical systems that allow us to reduce the behavior of physical systems in this way, there is no guarantee that all physical systems should be amenable to this trick, and that, on the contrary, there appears to be a very large class of systems that will never be so reduced. Running a simulation of such a system would then be the only viable way to find out how it behaves.

Can this idea of computational irreducibility be proved? Almost certainly not: at this stage it is an empirical observation about the nature of the world, and there may

never be any way to formally prove that a given system is permanently irreducible. It has the status of a meta-law of the universe: a law that says that not all the regularities in the universe can be described with equations that can be solved.

The main implication of computational irreducibility is that our faith in the power of mathematical descriptions of the physical world is misplaced: a system can show regularities in its overall behavior without there being any analytic explanation that connects those regularities to local mechanisms.

2. The Relationship Between Complex Systems and AI

One of the main arguments advanced in this paper is that complexity can be present in AI systems in a subtle way. This is in contrast to the widespread notion that the opposite is true: that those advocating the idea that intelligence involves complexity are trying to assert that intelligent behavior should be a floridly emergent property of systems in which there is no relationship whatsoever between the system components and the overall behavior.

While there may be some who advocate such an extreme-emergence agenda, that is certainly not what is proposed here. It is simply not true, in general, that complexity needs to make itself felt in a dramatic way. Specifically, what is claimed here is that complexity can be quiet and unobtrusive, while at the same time having a significant impact on the overall behavior of an intelligent system.

2.1.Conway's Game of Life

One way to see this quiet-but-significant effect is to look at a concrete example. This section is therefore centered on a set of variations on one of the most elementary of complex systems: John Horton Conway's cellular automaton called "Life" [5].

Life is a cellular automaton based on a 2-dimensional square grid of cells, with changes happening simultaneously to all of the cells at every tick of a global clock. Each cell can be either ON or OFF, and the rules for what the cell should do after the next tick depend only on its own state and the state of its eight nearest neighbors at the current time:

- If the cell is currently ON, stay ON if there are 2 or 3 neighbors ON, else go to the OFF state.
- If the cell is currently OFF, switch to the ON state if 3 neighbors are ON, else stay in the OFF state.

As is now well known, Life displays many patterns of activity that are stable in the sense that there is a period (a number of ticks) after which a pattern repeats exactly. There are also patterns that are not periodic, but which are salient for some reason. An entire zoo of "creatures" has been discovered in this miniature universe, including Gliders that move at a regular speed across the plane, Glider Guns that make Gliders, and Fuses that burn down like real-world fuses. These creatures are clearly an example of observed global regularities in the behavior of the system.

The most interesting question about Life concerns the explanatory relationship between the creatures and the local rules: is it possible to write down a formula that, for example, would predict all of the creatures that exist in Life, given only the definition of the local rules? The GLD says that this is not feasible: there is likely to be no analytic relationship between the local rules and the observed patterns.

2.2.Dynamic Complexity and Static Complexity

Conway's Life can be used to illustrate an important distinction between two different interpretations of how complexity manifests itself: what might be called "dynamic" regularities and "static" regularities. This distinction has to do with the fact that there is more than one level at which global regularities can appear in a complex system. The obvious level is where patterns of cell activation are noticed by us because they are stable. These are dynamic regularities because they occur in the moment-to-moment functioning of the system.

However, a second type of regularity can be seen in the search activity that Conway and his colleagues undertook when they were trying to find a good set of rules to define the game. Conway began with a desired global feature of his target system: broadly speaking, what he wanted was a high value for a parameter that we can refer to as the "generative capacity" (or "G") of the system—he wanted to have comparable numbers of ON cells being created and destroyed. According to one account, he found a set of rules with the desired global property,

> " ... only after the rejection of many patterns, triangular and hexagonal lattices as well as square ones, and of many other laws of birth and death, including the introduction of two and even three sexes. Acres of squared paper were covered, and he and his admiring entourage of graduate students shuffled poker chips, foreign coins, cowrie shells, Go stones or whatever came to hand, until there was a viable balance between life and death." [6]

Why did he not do some mathematical analysis and construct a function to predict the generative capacity from a given system design? Because such a function would have been extremely hard to derive, especially given the wide selection of grid types, gender valencies and rule structures under consideration. Complex systems theorists would now say that such an analytic function is practically impossible.

Although this "generative capacity" parameter is certainly a global regularity, it is a far less obvious type of regularity than the moving patterns of cell activation. To the extent that the generative capacity is not derivable from the local rules and architecture of the system, this is a bona fide example of complexity—but it is both static and more global than the dynamic complexity of the patterns. Static regularities are about characteristics of the system that are to some extent time-independent: whereas a dynamic pattern comes and goes, a parameter like the generative capacity has to be measured over a long period of time, and refers to the system and all of the patterns that occur in it. So there are two senses of "global" at work here: global structures that are larger than single cells, but which can be recognized in a modest amount of time, and global characteristics that encompass the behavior of the system and all of its patterns.

2.3.Searching for Static Regularities

This distinction is important because in some complex systems the dynamic regularities —the equivalent of the creatures in Life—may be the ones that grab all the limelight because of their salience, whereas there may be other, more subtle and almost invisible static regularities that turn out to be the most important features of the entire system.

We can see this point more graphically if we consider what might have happened if Conway had searched for systems characterized by something more ambitious than a high value for the generative capacity. At a very simple level, for example, he could have looked for systems that generate large numbers of interesting creatures, rather than just a good balance between cell creation and destruction. Note that this new

concept of "creature density" is a more subjective characteristic, but this subjectivity would not stop us from searching for systems with a high value for the creature density.

Would creature density be classified as a global regularity that is hard to predict from the local rules that govern the system? This is certainly possible, and empirical experience with large numbers of other complex systems indicates that it is likely. Is creature density less valid as a global regularity because it is more subjective? Not at all.

Now imagine a more ambitious goal: a system with a definite boundary, and with input patterns arriving at that boundary from outside, together with the requirement that the system should respond to each external pattern by producing an internal response that is unique in some way. For every pattern appearing at the boundary, one response pattern would have to be generated inside the system.

A further requirement might be that these response patterns be learned over time, and stand in a one-to-one correspondence with the input patterns, so that the internal response could be used to identify a previously seen input pattern.

At an even more ambitious level, consider systems that do all of the above, but which also seem to collect together entire classes of subjectively "similar" boundary patterns, in such a way that all patterns in the class trigger the same internal pattern. (The attentive reader will recognize that this has already been done: unsupervised neural nets can discover limited kinds of pattern in this way [7]).

Finally, imagine a system in which multiple external patterns are allowed to impinge simultaneously on different parts of the boundary, with patterns standing in various "relationships" to one another, and with the requirement that the system produce internal patterns that encode, not just classes of similar patterns, but meta-patterns that are "arrangements" of basic-level patterns (where an arrangement is a group of pattern-instances standing in a certain relationship to one another)—as well as patterns of these higher level patterns, and patterns of those second-level patterns, and so on, without limit.

We have clearly progressed well beyond Conway's modest goal of finding systems that maximize a global parameter called "generative capacity," so maybe it is time to invent a new term to parameterize the degree to which a system exhibits the particular, rather exotic type of global regularity described in the last couple of paragraphs. Call this parameter the "Lintelligence" of the system.

Just as it was for the case of the creature density parameter, we would expect this new parameter to be a global, static regularity, and we would also have no reason to treat it as a less meaningful concept just because it was subjectively defined.

Now imagine that we are looking at an example of a High-Lintelligence cellular automaton. What attracts our attention (perhaps) are the dynamic patterns of activated cells, and if we were interested in complexity we might be impressed by the way that these global structures could not be derived from the local mechanisms. However, if we are interested instead in the overall Lintelligence of the system, the local patterns are a distraction, because the Lintelligence is just as underivable from patterns as the patterns are from the substrate.

The name "Lintelligence" was designed to be suggestive, of course. Could it be that the thing we colloquially call "intelligence" is a global, static regularity like the one just described? Not as simple as Lintelligence, but perhaps an even more elaborate and exotic extension of it. This seems entirely reasonable—and if true, it would mean that our search for systems that exhibit intelligence is a search for systems that (a) have a characteristic that we may never be able to define precisely, and (b) exhibit a global-local-disconnect between intelligence and the local mechanisms that cause it.

2.4.Defining Intelligence

It might be worth noting one implication of this view of what intelligence is. If intelligence is a global, static regularity, we would not expect there to be any compact definition of it. This is consistent with the fact that people find it extremely hard to pin down what intelligence is. Common usage seems to indicate that intelligence is just the cooperative activity of a cluster of mechanisms, that there is no single choice of this set of mechanisms (so there could be many varieties of intelligence), and that there are graded degrees of intelligence, possibly forming a continuum all the way from the simplest thermostat to the human genius level and beyond.

The idea that intelligence could be reduced to a compact, non-circular definition in terms of "agents" and "goals," or that such a definition could be given a rigorous mathematical formalization, would then amount to nothing more than mathematization for its own sake: a *reductio ad absurdum* of the commonsense definition.

2.5.Multiple Levels

In the sequence of hypothetical systems just considered, there was an assumption that the dynamic patterns would always be obviously complex: that the whole system, from top to bottom, would look like a seething cauldron of emergence, in which nothing could be explained by anything going on at a lower level.

This is not the case: large systems can have structure at many levels of description, and some of those levels can seem more complex than others. In the case of the Life automaton, for example, it is possible to step up from the cell level and note that some of the discovered creatures are actually made out of conjunctions of other, smaller creatures. So there is a sense in which, at the creature level, objects can be used as building blocks for larger objects. What is interesting about this higher level of description is that it is more orderly—more mechanism-like and less complex-like—than the lower level, because the components can be used in a fairly predictable way to assemble larger structures.

In general, a complex system does not have to show its complexity at every level of description. There might be a level at which it gives a fair approximation of being a mechanism-like system. That "fair approximation" label is important, of course: any attempt to pretend that Conway's Life is not complex, and that it could be substituted with a facsimile system in which the lowest level was eliminated, leaving the level 2 objects as the base level, would almost certainly fail. In the case of Life we might try to build a facsimile system in which the basic units were creatures that were explicitly coded to interact in the ways that known Life creatures do, but although this system could very well be rigged so as to work like the original for a while, eventually its behavior would diverge from the real thing.

This question of multiple levels of description is relevant to AI because we try to build intelligent systems using ideas gleaned from insights about our own thought processes. We sometimes talk of the basic units of knowledge—concepts or symbols—as if they have little or no internal structure, and as if they exist at the base level of description of our system. This could be wrong: we could be looking at the equivalent of the second level in the Life automaton, therefore seeing nothing more than an approximation of how the real system works. (This argument is not new, of course: it was one of the main motivations cited by the early connectionists [8]).

2.6.Conclusion: Hidden Complexity

Combining the above notion of global, static regularity with the idea that there can be levels of description that are not obviously complex, it seems entirely plausible that the overall intelligence of an AI system could be a global, static regularity that is elusively dependent on complexity in the system, while at the same time there is a level of the system at which the "symbols" interact in ways that appear more mechanism-like than complex-like. If such were the case, and if we took pains to avoid situations in which the complex nature of the symbols would most likely manifest itself, we might stay convinced for a long time that intelligent systems are not significantly complex.

This matter has serious implications for the methodology of AI. The property of "being intelligent" might turn out to be either a mechanism-like property, or it might be a global, static complexity. If it is mechanism-like (in the way that "being fast" is a mechanism-like property of being a car) then all is well with the standard methodology of AI: we can try to find the low level components that need to be combined to yield intelligence. But if intelligence is a global, static complexity, then we may be in the same position as Conway when he was trying to find the local rules that would generate the global, static complexity that he sought. In the very worst case we might be forced to do exactly what he was obliged to do: large numbers of simulations to discover empirically what components need to be put together to make a system intelligent.

If our insights into the thinking process are showing us components that are not at the base level of our system—if the symbols or concepts that we appear to be manipulating are just an intermediate level of description of the system—then we may think we have identified the laws of thought, and that those laws of thought do not involve a significant amount of complexity, but we may be wrong.

At the very least, there is an important issue to be confronted here.

3. Making a Paradigm Choice

This argument is difficult to defend in detail, not because it is intrinsically weak, but because at the heart of the complexity idea is the fact that if the GLD is real, it will be almost impossible to say for sure whether intelligent systems really do involve significant complexity.

In order to definitively show that intelligence involves a significant amount of complexity, we may have no choice but to build many complete AI systems and gradually amass a database of example systems that all seem to have plausible local mechanisms. With this database in hand, we would then have to stand back and ask how successful these systems are, and whether we are seeing a good relationship between the changes we make to the local mechanisms (in order to fix whatever shortcomings we encounter) and the intelligence of the overall system. If, as part of this process, we make gradual progress and finally reach the goal of a full, general purpose AI system that functions in a completely autonomous way, then the story ends happily.

But the complex systems argument presented here says that this will not happen. When we look at the database of example systems, accumulated over the course of perhaps hundreds of years of work (at the present rate of system-building progress), we may find that, for some inexplicable reason, these systems never actually make it to full, autonomous intelligence. It might take a long time to get a database large enough to merit a statistically significant analysis, but eventually we might reach the empirical

conclusion that the overall intelligence of the systems does not bear a good relationship to the quality of the local mechanisms in each case.

If we ignore the Complex Systems Problem, this type of empirical effort may be the only way to come to a firm conclusion on the matter.

As an alternative to such a multi-century empirical program, the best we can do to decide whether complexity is important is to look for evidence that the ingredients of complexity are present. In other words we should stop asking for definitive, analytic proof that complexity is a problem (which the complex systems community claim is an impossible goal, if they are right about what complexity is), and instead look at the other way to define complexity: look at the ingredients and see if they are there.

This is fairly straightforward. All we need to do is observe that a symbol system in which (a) the symbols engage in massive numbers of interactions, with (b) extreme nonlinearity everywhere, and (c) with symbols being allowed to develop over time, with (d) copious interaction with an environment, is a system that possesses all the ingredients of complexity listed earlier. On this count, there are very strong grounds for suspicion.

We could also note a couple of pieces of circumstantial evidence. First, on those past occasions when AI researchers embraced the idea of complexity, as in the case of connectionism, they immediately made striking achievements in system performance: simple algorithms like backpropagation had some astonishing early successes [9][10]. Second, we can observe that the one place complexity would most likely show itself is in situations where powerful learning mechanisms are at work, creating new symbols and modifying old ones on the basis of real world input—and yet this is the one area where conventional AI systems have been most reluctant to tread.

3.1.Paradigm Arguments

The problem with citing suggestive or circumstantial evidence in favor of the idea that complexity is both present and playing a significant role in intelligence, is that this is easily dismissed by those who choose to be skeptical.

For this reason, there is little point making further attempts to argue the case: this is simply a paradigm issue, in Kuhn's classic sense of that term [15]. Deciding whether or not the problem is serious enough to merit a change of methodology is, ultimately, a personal decision that each AI researcher needs to make separately. It would be a waste of time to argue about definitive proof of the role of complexity, because the quicksilver nature of complex systems could always be used to cast doubt on such efforts.

In that spirit, the remainder of this paper moves on from the goal of arguing the case for complexity, and instead tries to sketch the first outline of an alternative approach to studying and building cognitive systems.

4. The Theoretical Psychology Approach

The new methodology that I propose is not about random exploration of different designs for a general, human-level AI, it is about collecting data on the global behavior of large numbers of systems, while at the same time remaining as agnostic as possible about the local mechanisms that might give rise to the global characteristics we desire. Rather than lock our sights on one particular approach—logical inference, bayesian nets, genetic algorithms, neural nets, or some hybrid combination of these name-brand

approaches—we should organize our work so that we can look at the behavior of large numbers of different approaches in a structured manner.

The second main ingredient of the approach is the use of what we already know about human cognition. For reasons that will be explained shortly, I believe that it is not possible to ignore human cognition when we do AI. This combination of a close relationship with the data of cognitive science, an empirical attitude to the evaluation of intelligent systems, and the lack of any commitment to directly explain human cognition, is what led to the choice of name for this approach: this is, in effect, a form of "theoretical psychology."

4.1.Generalized Connectionism

The roots of the theoretical psychology approach go back to early connectionism. When connectionist ideas first came to prominence, the core principle was about more than just using neuron-like processing units, it was about exploring the properties of parallel, distributed systems, to find out by empirical experiment what they could do. It was also about the "microstructure" of cognition—the idea that symbols need not be just tokens manipulated by an external processor, but could be processors themselves, or even distributed aspects of clusters of processors. This emphasis on open-minded exploration and the rejection of dogmas about what symbols ought to be like, is closely aligned with the approach described here.

Interestingly, as the connectionist movement matured, it started to restrict itself to the study of networks of neurally inspired units with mathematically tractable properties. This shift in emphasis was probably caused by models such as the Boltzmann machine [11] and backpropagation learning [10], in which the network was designed in such a way that mathematical analysis was capable of describing the global behavior.

But if the Complex Systems Problem is valid, this reliance on mathematical tractability would be a mistake, because it restricts the scope of the field to a very small part of the space of possible systems. There is simply no reason why the systems that show intelligent behavior must necessarily have global behaviors that are mathematically tractable (and therefore computationally reducible).

Rather than confine ourselves to systems that happen to have provable global properties, we should take a broad, empirical look at the properties of large numbers of systems, without regard to their tractability.

4.2.The Role of Intuition

If the only technique available to us were a completely random search of the space of possible cognitive systems, the task would be hopeless. The space to be searched is so large that if we had no other information about the target we might as well just try to randomly evolve an intelligent system, and expect to wait a very long time for results.

It is not the case, however, that we have no information about how an intelligence might function: we have some introspective knowledge of the workings of our own minds.

Introspection is what AI researchers have always used as the original source of their algorithms. Whether committed to human-inspired AI or to the normative, rational approach that eschews the need to build systems in a human-like manner, the ideas that are being formalized and implemented today have their roots in the introspections of past thinkers. Even the concept of logical, rational thought was an idea noticed by the

ancient Greek philosophers who looked inside themselves and wondered how it was that their thoughts could lead to valid conclusions about the world.

Introspection has had a bad reputation ever since the behaviorists tried to purge it from the psychological sciences, but in truth the aura of contagion that surrounds it is just an accident of the sociology of science. There is nothing wrong with it, if it is used as a source of inspiration for mechanisms that are then systematically explored and evaluated. For this reason, one of the defining features of the theoretical psychology approach is a search for ways to make the back-and-forth between introspective ideas and experimental tests of those ideas as fluid as possible.

The fact that we are cognitive systems ourselves is a good enough reason to hope that the starting points of our systematic investigations will be better than random.

There is another, less tangible, way that intuition can play a role. If we simulate large numbers of systems that differ by marginal changes in system parameters, we have some hope of observing regularities in the mapping between local mechanisms and global system behavior. Complex systems theory does not say that there will be no relationship whatsoever between the low level and the high level (it does not say that a small change in the low level mechanisms will always lead to an irregular, apparently random change in the high level behavior), it only says that there is no analytical relationship. It may well be that when we start to engage in systematic variations of candidate systems, we discover by inspection and observation that certain changes in low level rules cause lawful changes in the overall system behavior. If this happens, we may be able to converge on a workable design for an intelligent system in a surprisingly short amount of time.

The only way to find this out is to do some real science.

4.3.Neural Inspiration

If a grassroots form of connectionism is the way to go, would it also be a good idea to take the concept of neural inspiration more literally and only study systems that are as close as possible to the ones found in the brain?

It might be a mistake to impose this kind of restriction, because of the access to introspection mentioned above. At the conceptual level (as opposed to the neural level) we have some direct, introspective information about what is going on and why, but we have no such internal access to the workings of the mind at the individual neuron level. We simply cannot introspect any aspects of individual neurons, synaptic vesicles or dendrites, not least because we have little ability to report subjective events at the millisecond timescale.

And although we do have some external access to the functioning of the brain, through brain mapping techniques, signal interventions and post-mortem examination, our ability to get fine detail, and to link that detail to the cognitive level, is a subject of fierce debate within the cognitive science community [12]; [13].

4.4.Frameworks and Quasi-Complete Systems

What does it mean to engage in a program of "systematic exploration" of the space of cognitive systems? To be systematic, such a program needs to be unified by a common conceptual framework that is explicit enough that it allows the relationships between systems to be clearly seen.

The idea of a "framework" is that it defines a set of choices for the broad architectural features of the class of cognitive systems that it expresses. For every one

of the various mechanisms that we might anticipate being involved in a complete cognitive system, the framework should have something to say about how that mechanism is instantiated, and how it relates to the rest of the system. These specifications should be explicit enough that a complete system could be constructed in accordance with them.

It would also be vital to keep humans out of the loop as much as possible: systems need to be autonomous. If autonomy is not possible, the human involvement should be made explicit (and credibility discounted accordingly).

Any given research project within this framework would involve a particular instantiation of a system consistent with the framework, together with a particular focus on one or more components of that system. The idea would be to make systematic variations of some aspect of the design and look at the effect those changes had on the global system behavior.

Thus, someone determined to show that (for example) logical reasoning was a valid way to build intelligent systems would instantiate a set of traditional-looking symbols, inference mechanisms and all of the symbol-learning mechanisms and sensorimotor apparatus needed to connect the system with its environment. The mechanisms outside of the main focus of the research (all but the inference machinery, in this case) might be implemented in a simple, provisional way, but they would nevertheless be complete enough that the system could truly function by itself over a long period of time. If nothing at all could be done to build those other mechanisms in such a way that the system functioned at all, this would reflect badly on the logical reasoning mechanisms that are the focus of interest, but it would not invalidate them outright—that invalidation would only happen if, over the course of time, all attempts to improve the surrounding matrix of mechanisms ended in failure. (It is the feeling of this author that this eventual failure would probably happen, but the final arbiter of such a question would be empirical experiment).

More likely, frameworks would be used to explore a variety of non-traditional approaches to cognition, most of them in the generalized connectionist tradition. Whatever the philosophy that informs any given framework, however, the basic rules of the game would be to make both the framework and the systems instantiated within that framework complete enough to allow systematic variation of system parameters, and systematic comparison of the observed behavior of those systems.

This progression from general, loosely specified frameworks down to particular systems is a vital part of the process, and the framework part of the paradigm should not be dismissed as superfluous or scorned as unscientific. Anyone should be able to write down such a framework, for consideration by the community, so long as it is capable of being turned into a specific generator that yields instances of cognitive systems.

5. Conclusion

There is only space here to give a brief outline of the proposed theoretical psychology approach to AI, but even from this short account, it is clearly ambitious. One of its objectives is to persuade researchers to build and examine dozens of different types of system in a day, rather one type of system per career. Is this just a blue sky fantasy?

The way to make it possible is by means of a software development environment specifically designed to facilitate the building and testing of large numbers of similar,

parameterized cognitive systems. The author is currently working on such a tool, the details of which will be covered in a later paper.

More generally, any framework that purports to be the basis for a particular approach to AI—whether that framework is within the theoretical psychology paradigm or not—should be defined and represented by a software development environment that allows systems in that class to be constructed with relative ease.

What is needed, in fact, is not so much a software development environment as an SDE generator that enables anyone to build a customized tool for building their chosen class of cognitive systems. Such a generator would support a variety of frameworks, not just one.

5.1.Scruffs versus Neats

One way to summarize the philosophy behind this paper is to say that AI research has gone though two phases, and is about to enter a third. In their influential textbook of AI, first published in 1995, Stuart Russell and Peter Norvig point to a change in attitude in the AI community that occurred in the mid-1980s:

> Recent years have seen a sea change in both the content and the methodology of research in artificial intelligence. It is now more common to build on existing theories than to propose brand new ones, to base claims on rigorous theorems or hard experimental evidence rather than on intuition, and to show relevance to real-world applications rather than toy examples. Some have characterized this change as a victory of the **neats** – those who think that AI theories should be grounded in mathematical rigor – over the **scruffies** – those who would rather try out lots of ideas, write some programs, and then assess what seems to be working. Both approaches are important. A shift toward increased neatness implies that the field has reached a level of stability and maturity. (Whether that stability will be disrupted by a new scruffy idea is another question). [14]

Put this way, the distinction between neats and scruffies reflects very favorably on those who dominate AI research today.

The present paper disputes this analysis: the Scruffy era was defined by an engineering attitude to the problem, while the present day Neat era is defined by a mathematical attitude that gives a comforting illusion of rigor simply because of the halo effect caused by large amounts of proof and deduction. The transition from Scruffy to Neat looks more like the transition from Intuitionism to Behaviorism in psychology, and Neat AI has the same spurious aura of rigor that Behaviorism had.

No amount of mathematics can compensate for fundamental axioms that, when examined closely, turn out to be just as speculative as the ones that drove the engineers who came before. The complex systems perspective would argue that the overall performance of Neat AI systems will only be clear when complete examples of such systems—including all of the sensorimotor and learning mechanisms needed to ground them—are actually available for inspection and have been shown to converge on real intelligent behavior. Neat AI has scrupulously avoided such a showdown, so there is not yet any reason to believe that the assumptions at the root of the impressive-seeming mathematics are any more valid than the Scruffy assumptions that came before.

What we need right now is neither engineering nor mathematics, but science. We should be devising carefully controlled experiments to ask about the behavior of different kinds of systems, rather than exploring a few plausible systems chosen by instinct, or augmenting the same kind of instinctually-chosen systems with mathematics as a way to make them seem less arbitrary and more rigorous. Both of those old approaches involve assumptions about the relationship between the high-level

functionality of AI systems and their low-level mechanisms which, from the point if view of the Complex Systems Problem, are untenable.

References

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

[2] Holland, J. H. (1998) “Emergence.” Helix Books, Reading, MA.

[3]Horgan, J. (1995) “From complexity to perplexity.” Scientific American 272(6): 104-109.

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

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

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

[7] Kohonen, T. (1987) “Self-organization and associative memory.” Springer: Berlin.

[8] 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.

[9] 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.

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

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

[12] Harley, T. A. (2004). “Promises, promises. Reply to commentators.” Cognitive Neuropsychology, 21,51-56.

[13] Harley, T. A. (2004). “Does cognitive neuropsychology have a future?” Cognitive Neuropsychology, 21, 3-16.

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

[15] Kuhn, T.S. (1962) “The structure of scientific revolutions.” University of Chicago Press, Chicago, IL.