Towards a Theory of Open-Ended Evolution

Paul Chiusano
University of Michigan
pchiusan@umich.edu

Working Paper
12/30/2004

Abstract

This paper is an effort to begin serious work toward a theory of open-ended evolution (OEE). We begin by defining OEE to mean *continued innovation.* With this intuitive definition in mind, we closely examine metrics for characterizing behavior of evolutionary systems and argue that none are sufficient for quantifying open-endedness, but there are still avenues to explore. Working from our relatively simple definition, we are nonetheless able to reduce our root requirement of continued innovation into three, slightly more specific requirements, summarized as: 1) *potential for innovation* 2) *support for innovation* and 3) "*reachability" of innovation.* We then discuss how these requirements might be broken down further. In particular, we take up the question of whether the biosphere faces qualitatively different challenges in meeting these requirements than do systems like technological evolution. Our analysis seems to indicate that these seemingly dissimilar systems face identical challenges in meeting our OEE requirements (albeit on different scales), and we use this idea to aid us in attacking the third requirement—reachability—which seems to be the most challenging to ensure. We argue that two attributes: coevolutionary parts-sharing, and a "helpfully constraining" development process, may be important for establishing reachability.

Contents

1. Visions of a Theory

In the past fifteen years, cheaply available computers have become powerful enough to operationalize, using simulation, almost a century and a half of evolutionary theory. The results have been disappointing. Artificial life (ALife) and its sister field of evolutionary computation (EC) have so far failed to produce systems or models rivaling the biosphere's evolution in terms of creativity, adaptability, elegance, or most importantly—*open-endedness*. Though some have casually suggested that artificial evolutionary systems are still bounded by computational power, there is widespread agreement in these fields (and, I believe, strong evidence) that our understanding of evolution is somehow lacking. Simulation often makes for a good litmus test of the concreteness of any theory: if we are unable to simulate some phenomenon, there is a good chance we don't actually understand how it works.

"Open-ended evolution," as ill-defined as it is, remains a useful guiding concept for researchers in ALife and EC—building a model or system that exhibits this property is approaching holy-grail status in these communities. Later in this paper, we will try to come to a more precise definition, but for a start, we will consider an evolutionary system to be open-ended if it exhibits *continued innovation.*

This definition sidesteps the controversy surrounding the idea of "progress." Evolutionary biology has (rightly so) always been cagey on the notion of "progress"—whether it occurs, what it is, if it is inevitable, and so on. The very word "progress" seems to imply that evolution has some sort of definite goal or direction; as Dawkins warns in [12], "evolution has no long-term goal. There is no long-distance target, no final perfection to serve as a criterion for selection… In real life, the criterion for selection is always short-term, either simple survival or, more generally, reproductive success" (pg 50). Likewise Maynard Smith and Szathmáry spend a good chunk of their preamble in [27] warning us of the so-called "fallacy of progress." They note: "There is no reason to regard the unique transitions as the inevitable result of some general law: one can imagine that life might have got stuck at the prokaryote or at the protest stage of evolution" (pg 3).

And yet, here we are. The biosphere is evidently capable of spectacular continued innovation, and it is still a mystery how we might go about reproducing this phenomenon in our artificial systems. Researchers in ALife or EC—I think as a result of working with cold, hard simulation—have become pragmatic about these issues: most would probably be satisfied building a system or model which merely held the *potential* to evolve in a way that was as creative and innovative as the biosphere. Although it would be interesting to examine whether potentially open-ended evolutionary systems regularly achieve their potential on a particular 'run of the tape', we'd likely settle for a model that achieved this potential on *any* run.

Innovation also is perhaps something we all can agree on. As we will see, formally capturing what we mean by innovation is difficult, but most would probably agree that the biosphere is a system that has produced continued innovation, while artificial systems to date have not. The controversy is whether those innovations follow any particular long-term trend, whether that trend was inevitable, and whether another run of the tape or some other evolutionary system would experience a similar trend. In this paper we will focus on OEE as continued innovation, and punt on the issue of whether these innovations follow any particular long term trends. There may well be long-term trends in the biosphere's evolution that are not explained by the fact that any innovating process has a starting point, but I doubt whether these trends would be the same if we were talking about other, perfectly valid types of evolution, like the evolution of technology or scientific understanding.

It is extremely unlikely that every single aspect of the biosphere and the physical laws that govern it are relevant to its apparent open-ended evolution. More likely, only some aspects are relevant and—one can always hope—only a very few aspects are relevant and they are elegantly simple to boot. There are hints that such an outcome is possible, even probable. The fact that other evolving systems, like human technology or scientific understanding, also seem to be open-ended (but in very different ways), leads me to believe that open-endedness is a general property of certain complex evolving systems. It may therefore be possible to formulate a useful theory of open-ended evolution with extremely wide applicability.

To date, ALife and EC have focused a large proportion of their resources on what might be called "exploratory" research, which detractors are quick to dismiss as unscientific and undisciplined drivel. Actually, work like this can be very important for scientific progress. But what has become clear is that building a model exhibiting "open-ended evolution" is much more difficult than anticipated. The repeated failure of these models suggests that there are deeper questions here—if we take some time now to try to figure out more precisely what these questions are, we will have a much clearer idea about how to proceed. Thus, I don't see ad hoc model-building as the best way to move forward right now, and especially not models that "lack explicit theoretical grounding" [41]. Even if an ad hoc ALife model did manage to demonstrate "open-ended evolution," (however that term is defined) where would that leave us? We would still be unsure, for example, exactly which attributes of the model were needed and which weren't; we would be unsure why certain attributes lead to so-called "open-ended evolution" (whatever that meant, exactly), and we would have no disciplined way of mapping any of our OEE requirements or explanations to other evolutionary systems.

There seems to be some consensus in the ALife community that 1) understanding open-ended evolution is of the highest priority and 2) it is time to move forward in this area. At the 2000 ALife conference, for instance, Bedau et. al. presented a list of "open questions," many of which deal either directly or indirectly with open-ended evolution [4]. Rasmussen et al. also report on a survey of ALifers at that same conference in [36]: of those surveyed, the most commonly cited critical failures of ALife are lack or rigor and theory, and the most commonly offered issues or questions that it is important for ALife to address are 1) evolution (specifically, open-ended evolution) and 2) ALife's theoretical foundations.

At the risk of naval gazing, this paper does not present any experiments or show any results —I argue we don't know enough at the moment to even know *what experiments to run.* There is a lot to discuss, and the discussion requires more space than the page or two introduction to a results-oriented paper. A bit of thought, discussion and analysis now can potentially save a lot of pointless, dead-end experimenting and model-building down the line.

1.1 Requirements

What would useful theory on open-ended evolution "look like"?

Well, it would consist of at least three parts:

1) A definition of open-ended evolution
2) A set of unambiguous sufficient minimum requirements which guarantee that a system is capable of open-ended evolution
3) An understanding of how and why these requirements lead to open-ended evolution

In order for the theory to actually be useful, the definition of 1) would capture and clarify our intuitive notions of open-ended evolution and allow us to classify evolutionary systems as open-ended or not with a high degree of certainty.[1] The requirements of 2) would be specific enough that a person understanding them could *trivially* implement a model exhibiting open-ended evolution as defined by 1). The understanding provided by 3) of the theory would lead to true and testable predictions about classes of evolutionary systems (i.e., that adding a particular type of interaction *X* to some system will lead to behavior *Y)*.

There are, I think, good reasons to impose the above constraints on our theory. In an effort to make study of OEE more rigorous, it is tempting to rush out and attempt a mathematically precise definition or metric for open-ended evolution. But we may not know enough right now to reduce OEE to such a definition: if we are not careful, we might end up—using a rigorous, mathematically precise framework—defining a different phenomenon than the one we are actually interested in. The second constraint—requirements that are actually implementable—is intended to prevent any hand-waving about what attributes enable evolutionary open-endedness. So, for instance, saying that one of the minimum requirements for open-ended evolution is to "ensure that simple genotypes are connected along variational pathways to 'complex' genotypes" is no good because it doesn't tell us anything about exactly *how* we might go about ensuring this (nor does it define what is meant by "complex"). If a theory doesn't tell me as a programmer what to implement if I want to observe the phenomena, the requirements need to be more specific. Of course, the ability to simulate something does not guarantee understanding; thus requirement 3). Regarding our final constraint, an "understanding" of how and why the biosphere exhibits open-ended evolution which made no testable predictions would not be much of an understanding, and probably wouldn't tell us anything about other evolutionary systems.

1.2 On Scientific Methodology

Note that I see the above formulation as an *end goal* for a scientific understanding of open-ended evolution. I see no problem in relying temporarily on a "working definition" or on vaguely specified minimum requirements, or on an incomplete picture of how the biosphere or some other system exhibits open-ended evolution. In fact, I see an iterative development

[1] Open-ended evolution probably is not an all or nothing phenomenon. In this case, replace 1) with a way of characterizing the *open-endedness* of an evolutionary system.

process as the only way to move forward. There is no reason why such a process cannot eventually converge on rigorous theory.

It may seem odd developing a theory of some phenomenon concurrent to efforts to define that same phenomenon. How can we begin searching for the requirements of OEE before we have even defined, precisely, what we are searching for? It is often rather casually stated that theory on OEE has not moved forward much as a result of failure to define the problem more precisely. The trouble with this thinking is that it assumes that our theory requirements are all-or-nothing. They aren't. A definition has degrees of precision, and we need not have the definition nailed down to mathematical formulas to begin an analysis of what the requirements of OEE might be. Using a definition that is not much more precise than the one already given, I believe it is possible to make significant progress toward illuminating what the real issues are. Eventually, of course, we will need to revisit our working definition and make it more precise, but a little vagueness now does not prevent us from moving forward.

Furthermore, *all* of the requirements of our imagined theory are strongly interdependent. A more rigorous definition of open-ended evolution would benefit from having a number of systems to study, both natural and artificial, that capture our intuitive notions of the concept. At the same time, a clear definition of open-ended evolution could reduce developing an artificial model of it (which would contribute to parts 2 and 3) to mere problem solving. Likewise, a better understanding of how the biosphere (or some other system) apparently exhibits "open-ended evolution" could lead to a clearer definition of that term and insights into its minimum requirements.

There is a useful connection here to software engineering, which deals regularly with interdependent requirements. As researchers, we find ourselves in a situation similar to software developers attempting to build vaguely-specified software for a new client. Like the software developers, we are unsure of the requirements (we have no rigorous definition of open-ended evolution) and we don't know how to develop software which meets these requirements (we are unable to synthesize systems which exhibit open-ended evolution).[2] As in software engineering, the same drawbacks apply to the various approaches one might employ. The monolithic, classical software engineering approach would have the client and developers begin by exhaustively specify the project—this might be equivalent here to devoting all resources to originating precise metrics for quantifying open-ended evolution. Unfortunately, just as the output of the classical software engineering process runs the risk of not being what the client "really" wanted, a system exhibiting "open-ended evolution" according to well-specified metrics might not be manifesting the type of behavior that researchers were attempting to capture when they developed the metrics in the first place. I argue that we have already seen one such "iteration" of the classical approach, with Bedau et. al's evolutionary activity metrics [5, 6] providing the specification of open-ended evolution, and Maley and Channon's models providing the initial implementations [9, 26]. The other end of the spectrum—the "coding cowboy" approach—would call for little or no planning, and would devote nearly all resources to programming, or in this case, model-building.

[2] These points are debatable. Some might argue that Bedau's evolutionary activity metrics provide a definition of open-ended evolution, and Channon claims (using these metrics) that his 'Geb' system exhibits open-ended evolution. We will examine these claims later in this paper.

This paper advocates a strategy somewhere in between these two extremes, perhaps something akin to an iterative development strategy like Extreme Programming [2]. Along these lines, we will speak in terms of a *requirements tree*, rooted with a requirement "continued innovation." The process of nailing down the requirements of OEE—part 2 of our theory —is really nothing more than expanding the nodes of our requirements tree all the way to its leaves, where a "leaf" requirement is one specific enough to be implemented. Over the course of this lengthy process, we will reach nodes in the tree and realize that we need to refine nodes that we have already visited. Currently, for instance, in order to expand our requirement of "continued innovation," (and to examine possible candidate OEE metrics), we will need to have a more precise idea of what sort of innovation we are interested in. Later, after expanding more of the nodes of our tree, we may find that our notion of innovation needs to be refined further, and so on. This iterative process will continue until we converge on a precise OEE definition, and on OEE requirements that are specific enough to implement.[3]

The rest of this paper is organized as follows: we first synthesize a "working definition" of open-ended evolution as *continued functional innovation.* With this in mind, we examine metrics that have been developed for characterizing the behavior of evolutionary systems, with the hope that one of these metrics could serve to quantify open-endedness. After some analysis, we establish that this is likely not the case and identify more precisely some of the open issues which need to be addressed if any of the metrics for characterizing open-endedness are to be improved. From our relatively simple definition, we are nonetheless able to reduce our root requirement of continued innovation into three, slightly more specific requirements, which can be summarized as: 1) *potential for innovation* 2) *support for innovation* and 3) "*reachability" of innovation.* We then discuss how these requirements might be broken down further. In particular, we take up the question of whether the biosphere faces qualitatively different challenges in meeting these requirements than do systems like technological evolution. Our analysis seems to indicate that these seemingly dissimilar systems face identical challenges in meeting our OEE requirements (albeit on different scales), and we use this idea to aid us in attacking the third requirement—reachability—which seems to be the most challenging to ensure. We argue that two attributes: coevolutionary parts-sharing, and a "helpfully-constraining" development process, may be important for establishing reachability.

2. Open-Ended Evolution: A Working Definition

No generally agreed upon formal definition of open-ended evolution exists. Unlike, say, the mathematical construct of a *group,* open-ended evolution is not merely a label applied to a fixed set of properties. It is an attempt to characterize the behavior of certain dynamical systems, and dynamical systems do not typically have nice clean lines separating one class from another. For this reason, it seems more reasonable to speak of the *open-endedness* of a

[3] Upon "convergence" of this strategy, we will likely have a very good understanding of why our requirements (part 3 of our imagined theory) lead to open-ended evolution. Although there is always the chance that we could get lucky and stumble upon a set of sufficient OEE requirements, the approach here will basically require that we understand, at each point, why each set of child requirement nodes is sufficient to satisfy the parent requirement.

system. Along these lines, we might say that a GP system evolving the solution to the quintic problem has a very low degree of open-endedness. Ray's Tierra model is slightly more open-ended [37]. The biosphere and the evolution of technology [22, 23] both have a high degree of open-endedness, but these too, are probably ultimately bounded. Although going from our present artificial models to the biosphere represents a large jump in open-endedness, it's doubtful whether a discontinuity exists.

A formal definition of open-endedness will probably take the form of a metric. There are already a number of metrics for characterizing evolutionary systems; perhaps with some modifications, one of these metrics could serve to quantify open-endedness.

At this point, we should try to say more precisely what we mean by "innovation," since all evolutionary metrics can be thought of as making some assumptions about what is meant by this term, and we will need to evaluate whether these metrics are at least conceptually in the right ballpark. I have no rigorous definition, but I think I can safely say that innovation is defined at the level of *function*. For instance, when we speak colloquially about how much more "sophisticated" organisms or technology have become over evolutionary time, we are referring to functional attributes—homo sapiens are capable of *doing* things that other, "simpler" life forms can't do; computers today are capable of performing computations many orders of magnitude more quickly than computers fifty years ago. Thus, with any sort of evolutionary metric, we should ask ourselves whether it can tell us anything about the crucial, functional level.

Developing a general metric for characterizing any phenomena at the functional level is a difficult task. How humans recognize, say, functional sophistication is rather mysterious—we seem to rely on pattern recognition, abduction of purposes and goals, and other types of reasoning that would probably be difficult to reduce to mathematical formulas. Of course, we can always develop a metric to measure how well an entity is performing some particular function—EC systems do just this when they define a fitness function. But metrics for quantifying open-endedness need to transcend the particulars of specific EC or ALife systems, and it is legitimate to wonder whether such a metric is even possible.

Our only hope, it seems, is to develop measures of attributes not at the functional level, and to then establish that some configuration of these metrics is sufficient to guarantee the phenomena of interest at the functional level. Are there any potential evolutionary metrics that could succeed in making the leap?

2.1 Complexity Measures

One popular and intuitively appealing way of characterizing open-endedness considers OEE to be synonymous with unbounded growth in complexity. Of course, "complexity" could refer to any number of things, and so researchers have invested effort in defining more precisely what is meant by "complexity." Attempts to quantify complexity, like those discussed in [1, 24, 40] often rely on information theoretic approaches.

Unfortunately, we run into problems when we try to use a metric based in information theory to capture some phenomena at the level of function. Information theory deals fundamentally with representation, and one might say, deliberately ignores the function, purpose, or meaning of what is being represented. In other words, it makes no authoritative statements about function, and as we have argued, in the case of open-ended evolution, it is function that we are interested in. The biosphere's evolution is interesting because of what more advanced organisms are capable of *doing;* the fact that representations of their genotypes or phenotypes (might) contain more information is not really relevant.[4] Although it is sometimes the case that a functionally sophisticated process implies a correspondingly complicated representation, this isn't *always* the case, and the converse certainly isn't true. One can easily imagine a completely nonfunctional organism with a complex genotype, or a program whose source code is complex but which spits out jibberish. Likewise, a process which produced progressively more "complex" organism genotypes could be making absolutely no evolutionary progress in terms of function.

A similar argument goes for other attempts to clarify OEE in terms of concepts such as "hierarchical object complexity," and really any metric which deals fundamentally with representation (see for example [19]). These approaches will always be limited by the fact that a "complex" (however this is defined) representation does not necessarily guarantee "complex" functionality.

2.2 Bedau et. al Evolutionary Activity Statistics

Another, completely different possible approach to quantifying evolutionary open-endedness is Bedau and colleagues' evolutionary activity statistics [6]. Here again, we should ask ourselves: is this general style of metrics capable, in principle, of telling us anything about the level of function?

Before looking in detail at these statistics, let us try to refine what we mean by "functional innovation." This is actually a very difficult question—what makes a capability truly novel? Once again, though, we do not need to make our concept of functional innovation completely precise. We can make some progress merely by stating what we *don't* consider functional innovation. We will not consider an adaptation to be novel if it is merely a different representation (or a different way of encoding) functionality that the system has already discovered. Furthermore, we will not consider an adaptation to be novel if it is merely an "improved" version of some already existing capability. Although we haven't made precise the distinction between "new" and "improved," good examples of improved might be adaptations like: *sharper vision, faster reproduction rate, stronger bones,* and good examples of "new" might be things like: *the origin of vision, sexual reproduction, the vertebrae body-plan.* From these two concerns we can extract the following caveats:

1. There are often a large or even infinite number of ways of representing the same or highly similar functions.

[4] High "representational" complexity isn't even guaranteed for functionally complex entities—the often stated aphorism of Complex Systems is that simple rules can produce highly complex behavior.

2. There are often a large or even infinite number of possible functions within the same niche or paradigm.

Bedau et. al's statistics rely on three measurements: diversity, mean evolutionary activity, and new evolutionary activity. None of these directly measures anything at the functional level, but we hope that some configuration of these statistics, possibly the so-called "class 3" evolutionary dynamics, are sufficient to guarantee that the system is continually evolving functionally new entities. The signature of class 3 dynamics is: positive new evolutionary activity, unbounded diversity, and bounded (or unbounded) mean activity.

At first glance, it seems as though these statistics, measured in terms of "components" at the level of representation, would suffer from the same troubles as the complexity measurements we just examined. But the method for computing new evolutionary activity cleverly normalizes relative to a "neutral shadow," which replicates the real run in all respects except that selection is random. Roughly speaking, the logic is that if a newly created component is capable of surviving longer than a random component, chances are good that the component is adaptively significant. So even though these statistics are measured in terms of components, which are defined at the level of representation, the normalization relative to a neutral shadow seems like it sidesteps our first caveat—that there may be many representations of the same function, and that introducing a new representation does not alone guarantee that that new representation is not *identical in function* to some other representation already present.

So far so good. What about our second caveat? Although there are some issues with the neutral shadow normalization[5], let us suppose for the sake of argument that this metric or some variant can be used to reliably detect new evolutionary activity. New evolutionary activity, measured at the level of entity representation, is by itself insufficient to guarantee that the new evolutionary activity ever spills out into truly new functional paradigms, since a paradigm might be effectively "infinite." As an example, consider a system like the Evita model [5], in which a possibly lengthy entity representation codes for only a single functional attribute: reproduction rate. In such a system, a mutation that increases an entity's reproduction rate could very well prove to be adaptively significant. But even if this ratcheting up of reproduction rate continued indefinitely, we would not say that the system is exhibiting open-ended evolution because all adaptations fall within the same functional paradigm. No new capabilities are evolving, just greater and greater reproduction rates.

However, class 3 evolutionary dynamics require, in addition to positive new evolutionary activity, either unbounded mean activity, or unbounded diversity and positive mean activity, or both. As Bedau et. al point out, the decision about what to define as the system's components is an important one that can even potentially alter the class of evolutionary dynamics observed.

Under certain component definitions, the measurements for mean activity and diversity would prevent the classification of the above system as unbounded. An example of what we

[5] Maley, for instance, points out that this strategy has trouble detecting adaptively significant variation in evolutionary arms races. He suggests instead measuring the rate of propagation of the adaptation relative to a neutral model. Even if the adaptation is quickly displaced by another, it is likely significant if it spreads through the population more quickly than neutral adaptations.

might want for a component definition is the number representing an organism's reproduction rate. A reproduction rate of, say '10' represents a different component than a reproduction rate of, say, '11' (we haven't defined how these values are interpreted exactly, but this isn't terribly important). Since all values for this component fall within the same functional niche, it seems that competitive exclusion would a) limit the number of components in the system at any one time (thus bounding diversity) and b) prevent any particular component from persisting indefinitely (thus preventing the accumulation of activity, bounding mean activity). Organisms with reproduction rates of X will eventually be replaced by organisms with reproduction rates greater than X, and so on. Thus it appears that, although a given functional niche may be unbounded, the evolutionary activity statistics will "recognize" this and avoid classifying the system as unbounded.

Unfortunately, if we define components a little differently—more specifically, if the component is defined at a finer granularity than the system's level of selection, then components can be retained even though they contribute to the same function. Competitive exclusion might be capable of preventing peaceful coexistence of organisms or species in the same functional niche. But below the level of selection, components which contribute to the same function can coexist. In these circumstances, same-function components can continue accumulating activity and contributing to mean activity. I see this is a serious problem, since we actually expect an open-ended evolutionary system to define new units of selection [27].

To give an example of how all these ideas might operate to effectively "fool" evolutionary activity statistics into classifying a system as unbounded, consider the following toy Evita-like model. Organisms have only one functional attribute: reproduction rate, and compete only for space in the environment. When an organism in the system reproduces, it places its newly created progeny at some other location in the environment. If that location is occupied by another organism, that organism is effectively killed and replaced by the newly created organism. This results in intrinsic adaptation [35] toward higher and higher reproduction rates, possibly *ad infinitum*.

Although we would like to avoid assigning a special evolutionary category to this simple positive-feedback system, if components are compositionally below the system's unit of selection (the whole organism), we might not be able to avoid it. Suppose the component of the system is a binary mathematic expression built from primitive operators like addition and multiplication, and from terminal nodes like the real numbers. Suppose that organisms build expression trees for representing their reproduction rate. Obviously, there are an infinite number of binary mathematical expressions, and the expression trees grown will tend to become larger and larger as adaptation increases what is considered a "good" reproduction rate. The evolutionary activity statistics could measure positive new activity (new expressions are continually introduced that make an organism's reproduction rate expression greater), unbounded mean (or median) activity (since expression primitives in surviving organisms will accumulate activity) and possibly unbounded diversity (since there will be a growing number of components in the system).

For these reasons, I am somewhat suspicious of Channon's claims that 'Geb' is exhibiting open-ended evolution [8, 9], since his components, defined at the level of Lindenmayer system production rules, are at a finer granularity than the unit of selection (many production rules might combine to create the entire organism). It does not help that

Channon himself has a hard time determining what the organisms in Geb are actually doing: "…organisms have proved difficult to analyze beyond the above, even at the behavioral level. All that can currently be said is that they share characteristics of the previous species but are different" [10].

In other evolving systems that seem to be open-ended, it seems there can never be an unending arms race to infinity because certain constraints limit infinite progress in any single functional paradigm. Perhaps this is part of the importance of things like conservation of mass and energy and the second law of thermodynamics: they impose constraints that keep any particular paradigm bounded and thus force evolution to find new paradigms if it is to continue. Other types of evolution do not have *exactly* these same constraints, but they do seem to have limitations in place that prevent the continued exploitation of the same strategy. Take for instance the evolution of software—faced with the growing need to write more complicated programs, people could not respond by simply writing longer and longer assembly programs *ad infinitum*—our limited mental capabilities kept the assembly language paradigm bounded and forced the development of higher level languages. As it turned out, work done on grammars for programming languages turned out to be useful for other disciplines like linguistics. This sort of *coevolutionary parts-sharing*, we will argue later, is extremely important for continued innovation.

The designers of artificial systems are free to make whatever decisions they want concerning the underlying laws of the artificial universe. They are free to ignore any or all of the constraints on evolution that inevitably crop up in real-world evolutionary systems. While these constraints might make it more difficult for evolution to get started, perhaps they are in fact necessary for sustained progress. What could be extremely useful is a body of theory on what sorts of constraints lead to "paradigm boundedness." If we could be assured, based on the rules underlying an evolving system, that all paradigms were bounded, then continued positive new evolutionary activity could be sufficient for us to declare that the system was exhibiting open-ended evolution.[6] Guaranteeing something like this might not be possible though for an ALife system—since ALife systems endogenously create new functional paradigms, we won't know what paradigms will exist in advance. Is it possible to establish that all possible paradigms that could be created are assured to be bounded when we don't know what they will be in advance?

The Trouble with Diversity

There is one last issue regarding diversity that we should address. In [26], Maley presents a series of toy models, two of which he classified as exhibiting unbounded evolutionary activity. In Urmodel 3, parasites evolved bit-signatures and selection was based on how well these matched these signatures were relative to an externally supplied "host" bit vector. In Urmodel 4, hosts coevolved their own bit vectors, with selection now exerting pressure on hosts to differentiate their signatures from those of parasites.

[6] This isn't quite true. We can't literally run artificial systems indefinitely. A finite paradigm could still be huge, and we may not be able to run the system long enough to see whether evolution actually moves on after that paradigm is exhausted.

In these models, the system is technically evolving into new niches and diversity is increasing, but we can see, looking down from above, that all these niches belong in the same category. The hosts or parasites will never do anything other than change their signatures. As Maley says, such a system "would never surprise us." Likewise for Channon's Geb model—to me it does not seem very open-ended if organisms must always select an action from {reproduce, fight, turn, move forward} and cannot evolve truly novel capabilities.

Perhaps there is a useful distinction to be made among different types of functions. A parasite's signature has a function which allows it to exploit certain hosts. But this sort of function seems qualitatively different than, say, the functionality afforded to us by opposable thumbs. One is merely the correct key for a lock—the other represents a truly novel capability. Aside from relying on visualization, intuition, and common sense, it is still an open question how can we know which of type of functional evolution we are measuring.

It *is* extremely important that we develop a precise way of characterizing open-endedness. As Shalizi says (regarding self-organization, but the arguments apply equally well here):

> The prevailing 'I know it when I see it' standard actually impedes scientific progress, since it prevents our developing even the rudiments of a theory of self-organization. Thus some researchers state that 'self-organizing' implies 'dissipative,' and others claim to exhibit reversible systems that self-organize, and no one can even say if they are both talking about the same idea. [39]

I would emphasize that we need to be careful about declaring, prematurely, that we have captured exactly what we mean by open-ended evolution. The fact that some metrics are aligned with our intuitive (and often very accurate) recognition of some phenomena in one or a few cases does not mean that the same is true for all cases. We should ideally evaluate metrics whether they can, *in principle*, ever make authoritative statements about the phenomena we are actually interested in—not based on whether they just happen to coincide with our intuition some of the time.

3. OEE Requirements

Although it appears we haven't yet reduced open-endedness to an ideal metric, the "working definition" provided in the previous section is good enough to make some useful progress toward uncovering its minimum requirements. Recall that we considered open-ended evolution to mean, at the very least, the continued production of entities with "new" (and not just improved) functional capabilities. Although we haven't formalized what we mean by "new," we can still use this definition to severely constrain what the minimum requirements of open-ended evolution could be.

In fact, I would argue that we can straightforwardly decompose our root requirement of continued functional innovation into the following three requirements:

1. The system must be capable, in principle, of producing and supporting novel verbs. (*potential for innovation)*
2. At all times during the life of the evolving system, there must always exist unfilled niches or unexploited functional paradigms. *(ongoing support for innovation)*
3. At all times during the life of the evolving system, at least some of these unfilled niches must be reachable from the system's present state. (*ongoing reachability of innovations)*

These requirements shouldn't be too controversial—saying that a system meeting these requirements will exhibit continued functional innovation is tautological. If a system supports novel verbs (1), and there are always innovations out there that would be supported (2), and always at least some of those innovations are reachable (3), then the system will continue to discover those innovations. By a "reachable" innovation, we mean one that can be accessed by applying the system's variational operators to some already existing entity.[7] Note that these requirements hold regardless of how we define innovation.

Can we go further? As we suggested earlier, we would like to continue expanding nodes of our requirements tree until reaching "leaf" requirements—requirements so simple and obvious that it is a trivial matter to build a model or system that satisfies them. We still have a ways to go. Up until this point, we have not really done anything other than clarify our goals and questions a bit. But now, in decomposing the above requirements further, my impression is that we are entering unknown territory.

Let's look at ways we might break down the above requirements, and try to tease out the questions involved in making such a breakdown more precise or authoritative. The first, somewhat obvious point is that if a system is going to continually produce entities with new functions (requirement 1), the system must actually have the potential to support many different sorts of functionality. If we know that the number of possible functions are few, then we know that the system cannot possibly continue to produce entities with new functional capabilities. Thus, we can rule out our Evita-like model because we can surmise that all individuals in that model will only ever reproduce at increasing rates. They will probably never develop generic response systems, evolve language and technology, and start writing papers about open-ended evolution. We do need to be careful with these sorts of sweeping appraisals—the Game of Life, one might argue only has the actions {spawn, persist, die}, and yet the system is capable of generating actions, like movement, which seem to transcend these primitives.

There is a question here, about whether the primitive verbs of the system can be somehow combined or strung together to effectively produce novel verbs at higher levels. Perhaps this question can be phrased in terms of computational completeness. One doesn't need many primitives to guarantee computational completeness, and I find at least plausible that even the highly complex behaviors of organisms can be represented as a sort of computation. Maley has established that Ray's Tierran assembly language is computationally complete [25], and Teller has established that GP, augmented with READ and WRITE operators, is as well [42]. Again we need to be careful here: just because a system's primitives are Turing-complete doesn't mean that the system is capable of continual production of new functions.

[7] Reachability is probably not an on-or-off statistic. Many states may be technically reachable, but the probability of reaching the state from any current entity might be vanishingly small.

Consider an evolutionary system where entities live or die based on how well they play Pong. Even if the language for expressing these entity's pong-playing logic is computationally complete, the palette of actual actions is fixed in advanced {move up, move down}. In other words, we run into problems when an infinitely expressive control logic is bottlenecked by a very limited fixed-in-advance set of verbs. Thus, one requirement of an open-ended system might be that the verbs available be fashioned from an expressive language, and that the evolving entities actually have access to this language.

Requirement 1 can be thought of as dealing with the system's underlying rules—whether they are capable, in principle, of supporting continued innovation. Requirement 2 is asking something of the system's selective pressures—the underlying rules may form a language capable of expressing an innovation like sexual reproduction, but this innovation actually needs to be supported as a useful capability by the system's natural selection.

Requirement 1 might fall naturally out of Turing-completeness; requirement 2 may be satisfied simply by ongoing environmental change, achieved by allowing evolving entities to interact in a shared environment. Arguments like this are common in ALife; Bedau summarizes nicely in [3]:

> Each organism's environment consists to a large degree of its interactions with other organisms. So, if one organism evolves an innovative adaptive behavior, this changes the environment of neighboring organisms. This environmental change in turn causes neighboring organisms to evolve their own new adaptive behaviors, and this finally changes the environment of the original organism. In this way an organism's adaptive evolution ultimately changes the environment of that very organism. The net effect is that the population's adaptive evolution continually drives its own further adaptation.

The above analysis is not meant to be authoritative, and a thorough treatment of requirements 1 and 2 (not possible here) above could be very helpful. For now, we will focus on requirement 3, which is more mysterious. Just because an innovation *would be supported* if it were to spring into existence does not mean that there is actually a viable evolutionary pathway to that innovation from any of the entities which are currently present. To give a few examples: a lab doing cancer research may have all the materials needed to synthesize a miracle cancer drug (1), and the lab would obviously be interested in developing such a drug (2), but will the lab actually synthesize the drug? (is the miracle drug reachable?) Likewise, in biology, we know that complex innovations like the eye are in principle possible given the laws of physics and chemistry (1), and we know that eyes are *supported* by biological natural selection because they are useful (2), but were eyes actually reachable? Apparently so, but why and how?

Practically speaking, it is often very easy to imagine *designing* more sophisticated functionality than what actually appears during evolutionary runs, which leads me to believe that the first two requirements are not the current bottleneck.[8] Taylor, for instance, created his 'Cosmos' system with the goal of modeling the evolution of multicellularity [41]. The primitives of the

[8] This does not in any way diminish the importance of thoroughly understanding how to satisfy the first two requirements. If in fact we are, without really intending it, already building systems that satisfy these requirements, we still need to know and understand exactly what we are doing.

system were defined in such a way that multicellular organisms, as well as sexual reproduction were possible (meaning a human could design organisms with these capabilities), however these capabilities did not evolve. Likewise, given a few hours, a human programmer could probably design a sophisticated and flourishing Tierran organism which would never arise naturally.

It is not enough that a system be merely capable of creating and/or supporting new functional innovations—these innovations actually need to be connected to other already existing entities along the system's variational pathways. If we imagine a graph where the vertices correspond to potential niches or paradigms and an edge between two vertices implies that we can reach one niche from the other, then continued innovation seems to require that this graph be connected, rather than a large collection of isolated finite islands.[9]

This isn't much of a revelation. But there is a lot going on during evolution and it's easy to get sidetracked and distracted by other issues that aren't this and are probably less important. Artificial life has championed the importance of so-called "intrinsic adaptation," (which I see as possibly satisfying requirement 2) where new niches are created endogenously as a result of adaptation, and yet the field seems to have ignored the question of how we can insure that these newly opened-up niches are actually reachable from any already existing ones.

Let's look at some more examples. In biology, the evolution of a stable genetic code opened up the potential paradigm of recombination as a viable means of exploring the genomic space. Was this paradigm actually reachable by evolution? The evolution of multicellular organisms, combined with a complex and dynamic environment, opened up the potential niche for organisms capable of higher reasoning. Was this niche actually reachable? In the non-biological realm, the invention of written language suddenly made the printing press a viable niche, though the printing press depended on a number of technologies that had not nor could not have been invented at the time when written language was first invented. Although we can now see that all these potential niches were in fact reachable by some valid evolutionary pathway, we don't seem to have any idea why they should have been. In other words, we can induce, based on many examples, that something like reachability must have held for biological evolution, but this doesn't actually tell us anything about *why* it should have, nor does it tell us what to do if we'd like our artificial system to have this same property.

Lenski et. al report in [24] on experiments with the Avida system in which they evolved organisms to compute progressively more complex Boolean functions. In their discussion, they note the following:

> Some readers might suggest that we 'stacked the deck' by studying the evolution of a complex feature that could be built on simpler functions that were also useful. However, that is precisely what evolutionary theory requires, and indeed, our

[9] This description glosses over some things—for instance, it seems that both the biosphere and the "technosphere" have actually altered what is connected in the niche network over the course of evolution. (take, for instance, the introduction of sexual reproduction in the biosphere or the discovery of computer simulation in technology) Also, when entities can evolve not just through variation and recombination but also composition, the image of a niche graph breaks down.

> experiments showed that the complex feature never evolved when simpler functions were not rewarded.

Evolution of complex features does indeed require that these features have precursors that are themselves useful. But here it is apparent that ALife/EC is working from a different direction than evolutionary biology—whereas evolutionary biology can just accept, based on repeated observations, that even the most 'complex' feature has useful precursors, ALife and EC actually need to *engineer* this into the system. If this property is to hold for an artificial system, it probably won't be by accident—we will need to put it there.

3.1 Is it worth looking outside biology?

We have speculated on ways that we might decompose our first two requirements for OEE, but the third requirement—*ongoing reachability of innovations*—seems more opaque. I would argue we don't actually know enough right now to break down this requirements further—we need more information in order to bring about the critical 'aha!'.

Now, it is interesting that we already know of at least one evolving system—the biosphere—which we know satisfies these requirements. And if the biosphere were a simple, abstract model with only a few components, tracking down what was responsible for its apparent open-ended evolution would be straightforward. Unfortunately, the biosphere is a complicated system and many of the laws that govern it may be either irrelevant to OEE, or they may be just one of many possible "implementations" of a more abstract requirement. Questions like: are conservation of mass and energy important for open-ended evolution? or Is a nontrivial genotype to phenotype mapping important for open-ended evolution? are thus difficult to answer. If we only have one system to examine—the biosphere—I see no principled way of determining what aspects to factor out.

Clearly it would be useful to have a working model of OEE in hand, but we have a bit of a chicken-and-egg problem here, alluded to earlier: without an understanding of how the biosphere exhibits OEE, we have a difficult time building a working model of the phenomena—and without a working model, or at least some other evolutionary system to compare the biosphere against, we have a difficult time determining what aspects of the biosphere are important for OEE.

But if an OEE system is simply one that continues to innovate, then open-ended evolution is fairly common. An obvious example is human technological evolution. Human tools (technology) have been evolving in an open-ended way, starting, we can presume, from simple tools like clubs and spears and continuing to the present day. Likewise we could probably consider the related evolution of scientific understanding to be open-ended, and probably also the evolution of art and music.

I am obviously not the first to point out that there are other systems besides the biosphere that can be understood in terms of evolution. Dawkins coined the term "meme" [14] to refer to refer to a unit of cultural transmission, and memetics has become an entire discipline in its own right. Dennet has championed the "meme's eye view" and argued at length that the principles of natural selection should not really be confined to just biology [15, 16]. But there is still plenty of controversy over whether ideas from evolutionary biology transfer (see for instance Orr and Dennet's exchange in Boston Review [17, 33, 34]), and a feeling that these other types of evolution, if they are "really" evolution, are in a different boat than biology. People, after all, are capable of doing things like "design" and "induction" and "inference," whereas biological evolution is "random."

But I think a closer examination of this intuitions reveals that technological evolution faces the exact same difficulties (albeit at a different scale) as biological evolution. We have argued that any evolving system, if it is to continue innovating, must satisfy our three requirements of 1) *potential for innovation,* 2) *support for innovation* and 3) *reachability of innovation*. The question is not whether genes and memes have the same characteristics, it is whether things like "design" give civilization a decisive upper hand in meeting these requirements—so decisive that it these sorts of cultural evolution belong in a fundamentally different category. Obviously, being able to do things like simulate an idea for an invention in one's head (or on a computer!) makes evolution go much quicker since it doesn't take millions of years to execute evolutionary experiments—but does it mean that civilization faces barriers that are qualitatively different from biology?

The answer seems to be 'no.' In [21], Kauffman discusses technological evolution in the context of rugged landscapes, and he makes the argument:

> [I]f Darwin proposed a blind watchmaker who tinkered without foreknowledge of the prospective significance of each mutation, I suspect that much of technological evolution results from tinkering with little real understanding ahead of time of the consequences. We think; biological evolution does not. But when problems are very hard, thinking may not help that much. We may all be relatively blind watchmakers. (pg 202)

Others have pointed to the limitations of human cognition and used it to justify "evolutionary" assumptions in modeling systems. Nelson and Winter, for example, in developing their evolutionary economic theory, note the following: [31]

> [T]he amount of information storage implicit in the successful continuation of the routinized performance of the organization as a whole may dwarf the capacity of the individual human memory. The complexity and scale of the productive process may far surpass what any 'chief engineer,' however skilled, could conceivably guide (pg 106).
>
> [T]he problem with the neoclassical metaphor…is not that it connotes purpose and intelligence, but that it also connotes sharp and objective definition of the range of alternatives confronted and knowledge about their properties… [it] ignores the fact that it is not at all clear *ex ante* what is the right thing to do (pg 250),

Although the ability to design gets us pretty far, it does not get us infinitely far, and compared to the whole of the adaptive space, it gets us almost nowhere. A person living in 500 BC had about as much of a chance of inventing a complex future artifact like the printing press as a collection self-replicating molecules in the primordial soup had of spontaneously developing into a paramecium. Likewise, someone alive in the 1500s could not have designed a modern computer, a fax machine, or an automobile. That artifacts like modern computers, airplanes, and skyscrapers actually have valid pathways back to the first tools is no less amazing than the idea that the human eye has a valid evolutionary pathway back to the first light-sensitive spot.

We might say that the *vision* of the technological evolution system is finite, just as in biology. Although human civilization can certainly see further from its current state than the biosphere, it cannot see infinitely far and so it is theoretically possible for it to get stuck. Although it has not happened (yet), I can imagine a situation where a civilization's technological progress grinds to a halt because no one is capable of seeing any useful new forms from the current state, just as Maynard Smith and Szathmáry can imagine that "life might have got stuck at the prokaryote or at the protist stage of evolution" [27] (pg 3). The mystery of why this has not happened yet is the essence of our reachability requirement.

Other claims about "great advantages" that human evolution of any kind has over the biosphere can I think be dealt with similarly. Although there are some problems with the "vision" metaphor,[10] it can still work as a useful parameterization of evolutionary systems. The biosphere has a very low vision, the technosphere slightly greater, but the two systems belong qualitatively in the same category, and we can expect that both face similar obstacles to continued innovation.

3.2 Ensuring Reachability

This section is intended to suggest possible avenues to explore in trying to further break down our reachability requirement.

The past ten years have seen a resurgence of anti-Darwinist arguments in the so-called Intelligent Design movement—at their core, many of these arguments can be understood as skepticism about whether certain complex evolutionary features actually have valid evolutionary pathways, that is—whether certain features are reachable. Looking at responses to these arguments could be useful, because they might say something about what it is about biology that makes even the most complex features reachable.

Michael Behe, a biochemistry professor at Lehigh University, argued in *Darwin's Black Box* that Darwinism was insufficient to explain much of biochemical complexity [7]. The book describes a number of biochemical systems (cell cilia and bacteria flagellum, the blood clotting cascade, cellular active transport, and the immune system) that embody what Behe calls "irreducible complexity." To roughly paraphrase, an irreducibly complex system is one that fails to function if any of its constituent components are removed. Behe's paradigmatic example is the mousetrap, which he says fails to function if any of its components are

[10] "Vision" makes one think in terms of visual range in three dimensional space. Obviously, three dimensional space is not a good analogy for the "adaptive space" of a complex evolutionary system. And clearly, the "range of vision" will depend on what direction the evolutionary system is looking. People are good at designing far into the future with some types of systems, terrible with others. And of course, vision is itself evolving—one could think of sexual reproduction as a "vision-enhancing" adaptation.

removed. According to Behe, such systems cannot evolve by Darwinian gradualism and must therefore have been designed.[11]

There are plenty of cogent responses to Behe, addressing both his general argument that irreducible complexity implies design and the specific examples of apparently insurmountable irreducible complexity described in the book (see, for instance, [11, 30, 32]). But perhaps the most interesting, and the one most relevant to the question of how to ensure reachability, is Miller's response in [29]. Miller makes the following argument: the machinery for performing some function *A,* could depend on a number of different components. It may be the case that all subsets of these components cannot form a useful, selectable machine for performing function *A,* but it is possible that they form a useful machine for performing function *B,* or *C* or *D…* Saying that a mousetrap requires all its parts to function *as a mousetrap* could be true (actually, even this has been contested: [28]), but it is much less likely that none of the parts could be put to useful work for some other purpose. Indeed, Miller has been known in lectures to break apart the five part spring-loaded mousetrap and demonstrate (amusingly) how various subsets of these components can be used for other purposes, including a tie clip, catapult, toothpick and nose ring [38].

This is more than just a cute demonstration—as Miller highlights in [29], the origins of complex, seemingly "irreducibly complex" biological machinery like the Krebs cycle or the bacterial flagellum can actually be explained using similar mechanisms. The evolutionary origins of the Krebs cycle were mysterious, but we now have strong evidence that "nearly all of the proteins of the complex cycle can serve different biochemical purposes within the cell" [29], and it is possible to construct highly plausible scenarios for the cycle's evolution.

A good name for this general phenomenon might be *coevolutionary parts-sharing.* It would be interesting and potentially useful to formalize ideas on parts-sharing and use them to further break down our reachability requirement. There are still plenty of questions here—how many other functions are also useful, and to what extent can machines for different purposes "share parts?" In the evolution of science, for instance, Hopfield adapted work done in statistical physics to build models of neural computation with content-addressable memory [20]. Likewise, Maynard Smith adapted ideas from game theory to explain certain types of cooperation that emerge in nature. This sort of interdisciplinary parts-sharing seems to be important, and if biologists were not allowed to read papers on game theory and computational neuroscientists were not allowed to read papers in statistical physics (if parts sharing were not supported by the scientific evolutionary system), the scientific advances above might not have occurred.

Organisms are less able to share functionality across species than people are able to share ideas across disciplines, but certainly within a species or an organism, there is significant potential for parts-sharing, as the Krebs cycle example demonstrates. Perhaps a succinct mathematical model could show how, given a broad range of assumptions for the number of functions being selected for, and the "shareability network" of the parts that compose these functions, the probability that even very complex bits of functionality have useful precursors

[11] Behe carefully shies away from identifying the designer: "Inferences to design do not require that we have a candidate for the role of designer. We can determine that a system was designed by examining the system itself, and we can hold the conviction of design much more strongly than a conviction about the identity of the designer" (pg 196).

becomes quite high. Paradoxically, it might be easier for evolution to evolve many different functions at once than it is to evolve just one or a few.

The Importance of a Development Process

It is looking as if the way to break down our reachability requirement is to ensure that a "large" proportion of the states that an evolutionary system is capable of producing be useful. When a lot of states are "useful," it becomes less and less likely that the system will ever find itself stuck on an island. We don't know what proportion of useful states is required—this seems like it would depend on properties of the "reachability network" underlying the evolutionary system—but it is still possible to speculate on ways that an evolutionary system might increase the proportion of useful states. The one that we have just examined—having many different sorts of functionality be evolutionarily useful—can be thought of as increasing this proportion of useful states by simply increasing the *absolute number* of useful potential states present in the system. If we are only interested in one function, and we are fashioning our entities from an expressive language, then the number of useful states is very small in comparison to the total. But if we are interested in many functions, and the machinery for one function can share parts with the machinery for another, then the number of useful states is much greater and it seems that the way these states are connected makes it significantly easier for even very complex bits of functionality to have useful precursors.

There is another way of increasing the proportion of useful states, and that is to *decrease* the number of possible states which are not useful. Dawkins has argued in [13] and [14] for the importance of a development process, and I think these arguments can now be seen in a new light—a development process drastically reduces the number of possible forms and (presumably) decreases the proportion of them that are not useful. We often consider biological evolution to be random, but I would argue that the constraints imposed by a lengthy development process from DNA to organism bring the "random" recombination and mutation closer to calculated design decisions. It is surprising that such a large proportion of recombination and mutation are nondestructive, and it could be very important for an open-ended evolutionary system to have a disciplined way of exploring the space of possible forms like what the development process of biology (and human design) provide.

Dawkins makes the argument that a development process is needed for continued adaptation. In a discussion about the evolution of complex animal organs, he makes the following point:

> [T]he ancestral organs did not literally change themselves into the descendant organs, like swords being beaten into ploughshares. Not only *did* they not. The point I want to make is that in most cases they *could* not. There is only a limited amount of change that can be achieved by direct transformation in the 'swords to ploughshares' manner. Really radical change can be achieved only by going 'back to the drawing board'… [14] (pg 260)

He continues:

> Maybe you can beat a sword into a ploughshare, but try 'beating' a propeller engine into a jet engine! You can't do it. You have to discard the propeller engine and go back to the drawing board (pg 260).

There are two points here: one is that there are limits as to what extent living things are capable of adaptation; the other, implicit point is that most alterations of a living organism will result in death.

This argument can be extended to all types of evolution. Whether it is biological organisms or technological artifacts, all "living" entities have some set of capabilities or functionality that have allowed themselves and their ancestors to persist and flourish in the evolutionary system in which they participate. A bird persists because it has (among other traits) the capability of flight which allows it to gather resources and energy from a physical environment. A pen persists by providing a human with the ability to write, which we apparently find useful. A speaker persists by being able to produce sounds when supplied with electrical current; a computer persists by being able to run programs and interface with human users. A scientific theory persists through its ability to explain, predict, and help people to understand phenomena. Although a pen that runs out of ink does not vanish into a black hole, and likewise for a broken computer, and we can still read about vitalism if we are curious, there is a sense in which these entities are *dead*—they no longer possess the capabilities that allowed them to come about and thrive in the evolutionary system in which they participate. Note that these metaphors are not intended as definitions of life or death; I am only using these terms to draw an analogy between the different types of evolutionary systems.

When Intel designed the Pentium 4 chip, they did not literally try rewiring the P3—even if this were possible, we might say that it would require killing the CPU in that it would remove its ability to perform computations, which is the whole reason why people keep CPUs around in the first place. Other human artifacts are more slightly more adaptable than CPUs, but ultimately suffer from the same rigidity. Large pieces of software are a good example—when a program reaches millions of lines of code, and understanding it in its entirety is impossible, really radical change also becomes impossible because doing so would require temporarily killing the program (removing the capabilities that made the program useful in the first place) and "reviving" the dead program would require superhuman mental capabilities. The fact that this happens in software, which is arguably more adaptable than any other type of human artifact, suggests that death and redevelopment could be essential for any evolutionary system that continues to innovate.

It seems like the vast majority of forms that are theoretically possible are actually incapable of supporting life. In order for rather stupid and limited processes like mutation, recombination, and design to actually get anywhere, the evolutionary system needs to constrain the space by working at a level where more of the states lead to entities capable of persistence. In technological and scientific development, this has taken the form of working at the level of ideas or designs, which are then developed or deployed; in biological evolution, it has taken the form of an elaborate process in which genotype develops into phenotype. Although there has been plenty of speculation on the importance of a development process from genotype and phenotype, I am not aware of any research that explains how such a process can work to reduce the number of potentially non-useful states. Based on our analysis so far, there is good reason to think that this capability of a development process could be quite important.

In general, it seems like not enough ALife and EC researchers give the development process the respect it deserves. Although there are a number of ALife (and EC [18]) systems that have a development process or some sort of genotype-phenotype mapping, there do not seem to be any disciplined grounding assumptions about how these processes might actually work to increase evolutionary potential (other than vague appeals to biology). Slapping together an ad hoc development process vaguely reminiscent of biological development is probably not the way forward here. But if the analysis presented here is correct, then we have at least have a better idea of what the goals or usefulness of such a process should be: a "helpfully constraining" development process would significantly reduce the number of non-useful states.

4. Conclusions

In this paper, we have argued for the importance of developing a theory of "open-ended evolution," outlined what such a theory might look like, and tried to take a few steps toward its realization.

We argued that the open-ended evolution label is appropriate for evolutionary systems that exhibit continued "innovation," and although did not formalize what we mean by innovation (this will need to be revisited soon), we said at least that we are interested in *functional* innovation—that is, the production of entities that are capable of doing "new" things. We showed that no current metrics for quantifying open-endedness quite manage to capture or ensure this trait and suggested some avenues to explore.

We then straightforwardly broke down our root requirement of continued functional innovation into the following three requirements:

1. The system must be capable, in principle, of producing and supporting novel verbs (*potential for innovations)*
2. At all times during the life of the evolving system, there must always exist unfilled niches or unexploited functional paradigms (*ongoing support for innovations)*
3. At all times during the life of the evolving system, at least some of these unfilled niches must be reachable from the system's present state (*ongoing reachability of innovations*).

We then speculated on ways we might break down these requirements. The first requirement might be answered by a system with underlying rules that give rise to a Turning-complete set of primitive verbs. The second requirement might be satisfied by a system undergoing "intrinsic adaptation." The third requirement might be satisfied by a combination of coevolutionary parts-sharing and a helpfully-constraining development process. All these arguments were somewhat speculative and we have tried to suggest ways they could be made more precise, or at least offered ideas for avenues to explore. We also argued that more information may be needed to break these requirements down further, and that this information might very well come from a closer examination of another open-ended evolving system like human technology or scientific understanding.

Thanks go to: Andrew Yinger, Mike Samples, Rick Riolo, Jason Daida, Blake Nicholson and many others.

5. References

[1] Adami, C., Ofria, C. Collier, T. Evolution of Biological Complexity. *Proceedings of the National Academy of Sciences*, *97*.

[2] Beck, K. *Extreme Programming Explained: Embrace Change*. Addison-Wesley Professional, Menlo Park, 1999.

[3] Bedau, M. Four Puzzles About Life. *Artificial Life*. 125-140.

[4] Bedau, M., McCaskill, John, Packard, Norman, Rasmussen, Steen Open Problems in Artificial Life. *Artificial Life*, *6*. 363-376.

[5] Bedau, M., Snyder, Emile, Brown, C. Titus, Packard, Norman, A Comparison of Evolutionary Activity in Artificial Evolving Systems and in the Biosphere. in *Fourth European Conference on Artificial Life*, (1997), MIT Press/Bradford Books.

[6] Bedau, M., Snyder, Emilie, Packard, Norman Classification of Long-Term Evolutionary Dynamics. *SFI Working Paper*.

[7] Behe, M. *Darwin's Black Box: The Biochemical Challenge to Evolution*. Touchstone, New York, 1994.

[8] Channon, A., Improving and Still Passing the ALife Test: Component-normalized Activity Statistcs Classify Geb as Unbounded. in *Proceedings of Eighth Annual Conference on Artificial Life*, (2003).

[9] Channon, A., Passing the ALife Test: Activity Statistics Classify Geb as Unbounded. in *Advances in Artificial Life: 6th European Conference*, (Prague, Czech Republic, 2001), Springer-Verlag Heidelberg.

[10] Channon, A. Toward the Evolution of Increasingly Complex Advantageous Behaviors. *International Journal of Systems Science*, *31* (7). 843-860.

[11] Coyne, J.A. God in the Details: The Biochemical Challenge to Evolution. *Nature*.

[12] Dawkins, R. *The Blind Watchmaker*. W. W. Norton & Company, New York, 1996.

[13] Dawkins, R. *The Extended Phenotype*. Oxford University Press, 1999.

[14] Dawkins, R. *The Selfish Gene*. Oxford University Press, 1989.

[15] Dennet, D. *Darwin's Dangerous Idea*. Touchstone, New York, 1995.

[16] Dennet, D. Memes and the Exploitation of Imagination. *Journal of Aesthetics and Art Criticism*, *48* (Spring 1990). 127-135.

[17] Dennet, D. The Scope of Natural Selection. *Boston Review* (October/November).

[18] Ferreira, C. *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*. Gepsoft, www.gene-expression-programming.com, 2002.

[19] Heylighten, F. The Growth of Structural and Functional Complexity During Evolution. in F. Heylighten, D.A. ed. *The Evolution of Complexity*, Kluwer Academic Publishers, 1996.

[20] Hopfield, J.J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Sciences of the USA*, *79*. 2554-2558.

[21] Kauffman, S. *At Home in the Universe.* Oxford University Press, New York, 1995.

[22] Kurzweil, R. *The Age of Spiritual Machines.* Penguin Books, New York, 1999.

[23] Kurzweil, R. The Law of Accelerating Returns, KurzweilAI.net, http://www.kurzweilai.net/articles/art0134.html, 2001.

[24] Lenski, R., Ofria, Charles, Pennock, Robert, Adami, Christoph The Evolutionary Origin of Complex Features. *Nature*, *423*.

[25] Maley, C.C. The Computational Completeness of Ray's Tierran Assembly Language. *Artificial Life*, *3*. 503-514.

[26] Maley, C.C., Four Steps Toward Open-Ended Evolution. in *Proceedings of the Genetic and Evolutionary Computation Conference*, (San Francisco, CA, 1999), Morgan Kaufmann Publishers, 1336-1343.

[27] Maynard Smith, J., Szathmáry, Eörs *The Major Transitions in Evolution*. Oxford University Press, Oxford, 1995.

[28] McDonald, J.H. A Reducibly Complex Mousetrap, http://www.millerandlevine.com/km/evol/design2/article.html, 2002.

[29] Miller, K.R. The Flagellum Unspun: The Collapse of "Irreducible Complexity". in Dembski, W.A., Ruse, Michael ed. *Debating Design: From Darwin to DNA*, Cambridge University Press, 2004.

[30] Miller, K.R. Review of Darwin's Black Box. *Creation / Evolution*, *16*. 36-40.

[31] Nelson, R., Winter, Sidney *An Evolutionary Theory of Economic Change.* Belknap Press, Cambridge, 1982.

[32] Orr, H.A. Darwin v. Intelligent Design (Again). *Boston Review.*

[33] Orr, H.A. Dennet's Strange Idea. *Boston Review* (June/July).

[34] Orr, H.A. Response to 'The Scope of Natural Selection'. *Boston Review* (November).

[35] Packard, N.H. Intrinsic Adaptation in a Simple Model of Evolution. *Artificial Life.*

[36] Rasmussen, S., Raven, Michael, Keating, Gordon, Bedau, Mark Collective Intelligence of the Artificial Life Community on Its Own Successes, Failures, and Future. *Artificial Life*, *9*. 207-235.

[37] Ray, T. An Approach to the Synthesis of Life. *Artificial Life*, *2*. 371-408.

[38] Saylor, F. Miller Continues Darwinian Defense at MIT *Science and Theology News*, June 2004.

[39] Shalizi, C., Shalizi, Kristina, Quantifying Self-Organization in Cyclic Cellular Automata. in *Proceedings of SPIE: Noise in Complex Systems and Stochastic Dynamics*, (2003).

[40] Standish, R. Open-Ended Artificial Evolution. *International Journal of Computational Intelligence and Applications*, *3* (167).

[41] Taylor, T. From Artificial Evolution to Artificial Life, University of Edinburgh, 1999.

[42] Teller, A., Turing Completeness in the Language of Genetic Programming with Indexed Memory. in *First International World Congress on Computational Intelligence*, (1994), IEEE Press, 136-146.