Essay 6

Artificial Darwinism: The Very Idea![*]

Barry McMullin

Dublin City University,
Dublin, Ireland.

[*] *This is a severely abridged version of selected material first presented in my Ph.D. Thesis (McMullin 1992e).*

Abstract

The realisation of artificial Darwinian evolution is one conceivable—indeed, more or less obvious—route toward the realisation of a growth of knowledge (or "complexity") in artificial systems. This paper explores the current state of the art in achieving Artificial Darwinism, and the prospects for further progress. In particular, I reassess the seminal work of von Neumann on evolution in cellular automata (von Neumann 1951; 1966a; 1966b). I also review the *Genetic Algorithm* (Holland 1975), and the `VENUS` (Rasmussen *et al.* 1990) and `Tierra` (Ray 1992) systems. I attempt to relate this to the work of Varela, and others, on the realisation of *autopoiesis* in related (discrete, 2-dimensional, homogenous) spaces (Varela *et al.* 1974; Zeleny 1977; Zelany & Pierre 1976), and I also revisit the Holland α-universes (Holland 1976; McMullin 1992d). I suggest that while both open-ended heredity (von Neumann style "self-reproduction") and spontaneous autopoiesis have been separately demonstrated in such systems, the combination of the two remains a difficult outstanding problem. I conclude by outlining an avenue for further investigation.

6.1 Introduction

There is a very large literature already in existence which bears on what I term *Artificial Darwinism*—i.e. the possible realisation of Darwinian evolution in artificial systems. Furthermore, work on this topic has recently received a new impetus with the (re?)emergence of the field now called *Artificial Life* (Langton 1989a; Langton *et al.* 1992; Varela & Bourgine 1992). The size and rapid growth of this literature precludes any attempt at a comprehensive survey or critique, and I do not pretend to provide one. Instead, this paper will be concerned with a selective review of work carried out by a small number of researchers. I shall concentrate particularly on von Neumann's seminal investigations, and I follow this with a discussion of what seems to me to be the most directly relevant subsequent work.

Von Neumann carried out his work in this area, for the most part, in the period 1948–53. He presented his ideas in various lectures over that period, and some limited discussion was also formally published around the same time (von Neumann 1951; Kemeny 1955). Von Neumann himself started work, in 1952–53, on a major book in this area, tentatively entitled *The Theory of Automata: Construction, Reproduction, Homogeneity.* However, he put this aside in late 1953 and, as a result of his untimely death in 1957, he was never to return to it. While the draft manuscript circulated fairly widely, it was only through the efforts of A.W. Burks that it was finally edited, completed, and posthumously published, together with a series of related lectures (also previously unpublished), under the general title *Theory of Self-Reproducing Automata* (Burks 1966b).

I contend that von Neumann's original work has been, at best, incompletely understood; and that the research programme which he proposed has foundered. Thus, a primary purpose here is to attempt a fresh evaluation and re-interpretation of von Neumann's work. In the light of this, I then go on to comment *critically* on the subsequent development of the field. My conclusion will be the unsurprising one that the problem of realising Artificial Darwinism, at least in the strong sense in which I am using that term, is extremely difficult; that progress in this direction has been very limited; and that any conceivable alternative strategies to realising this goal should be carefully explored.

6.2 Von Neumann's *Theory of Automata*

6.2.1 Von Neumann's Problem (P_v)

Although it seems to have been von Neumann's ultimate objective to formulate a single, comprehensive, and completely general, "theory of automata", I take the view that that objective has certainly not yet been achieved. Instead there exists a wide variety of more or less distinct "theories of automata", which are related in various ways, but which preserve their own unique characteristics also; and in what follows it will be necessary to consider at least a selection of these distinct theories. I therefore introduce some new terminology to facilitate this discussion.

I shall refer to some particular axiomatization of (abstract) automata as defining an Automata-System or *A-system.* Within the context of such a particular A-system I shall refer to the entities which are to be regarded as "automata" as *A-machines*. The set of all A-machines (with respect to a particular A-system) will be called the *A-set.* The possible "primitive" (irreducible) parts of an A-machine will be called *A-parts.* In general it must be possible to analyse the behaviour of any given A-machine in terms of its being composed of a number of A-parts, which are "legally" arranged or aggregated. I shall refer to an arbitrary aggregate of A-parts as an *A-*

structure.

Note that "A-structure" and "A-machine" are not, in general, synonymous, though they are clearly related. In fact, certain A-structures may not qualify as A-machines at all; and certain, distinct, A-structures may be regarded as instances of the "same" A-machine (in different "A-states")—i.e. an A-machine might be defined as some kind of equivalence class of A-structures. Indeed, it is conceivable that we could have two A-systems which incorporate exactly the same A-parts, and thus have exactly the same sets of A-structures, and yet which differ radically in their definitions of what constitutes an A-machine.

As well as this terminology specifically relating to automata, I shall also make occasional use below of a technical terminology regarding the abstract ideas underlying Darwinian evolution in general. The latter terminology is detailed in (McMullin 1992a), and I shall provide only a brief summary here.

Actors are individuals which reproduce, with some degree of heritability. A Similarity-lineage or *S-lineage* is a lineage of actors which includes, at each generation, *only* those offspring which are "similar" to their parent(s) in some specified way. Distinct, heritable, "similarities" (similarity-classes or *S-classes*) thus distinguish distinct S-lineages. In the general case, any given actor may be a member of many distinct S-lineages. In certain circumstances an S-lineage may grow consistently until limited by resource availability; and, in so doing, may exclude or eliminate one or more other S-lineages. This is S-lineage *selection*. *S-value* is a parameter of an S-lineage such that differences in S-value are predictive of the rate and ultimate outcome of selection. S-value corresponds to one of the common interpretations of "fitness" in evolutionary biology.

The birth of an actor with some heritable characteristic not possessed by any of its parents is called *S-creation*. S-creation initiates new S-lineages. If S-creation is *blind* or *unjustified* (in the sense of Campbell 1974a, 1974b) the actors are called Darwinian- or *D-actors*. A lineage of D-actors, incorporating multiple distinct S-lineages, whose evolution can be usefully described in terms of selection events between those S-lineages, is called a *D-lineage*. A system of D-actors, forming D-lineage(s), is called a *D-system*.

Some further terminology will be introduced below as the context demands. In particular, where it is necessary to restrict the discussion to some particular A-system, an appropriate subscript will be added, thus: A_X-system, A_X-structure, A_X-part etc.

Von Neumann's foundational problem in the theory of automata, which I shall denote P_v, was to formulate a particular A-system in such a way that the following distinct conditions are satisfied:

1. There should not be too many different "kinds" of A-part, nor should these be individually very "complex".

2. We require that some A-machines operate (in at least some circumstances or "environments") so as to acquire (somehow) further A-parts, and assemble them into new A-machines. A-machines of this sort will be called *A-constructors.* In general, we do not expect that all A-machines will be A-constructors, so that the set of A-constructors will be a proper subset of the A-set.

3. We require that some of the A-constructors be capable of constructing offspring which are "identical" to themselves.[1] We shall call these *A-reproducers.* A-reproducers may also, of course, be capable of constructing A-machines quite different from themselves. In general, we do not expect all A-constructors to be A-reproducers, so that the set of A-reproducers will be a proper subset of the set of A-constructors.

4. We require that there should exist some mechanism(s) whereby an A-machine can "spontaneously" change into a different, distinct, A-machine; these changes will be called *A-mutations.* We require that A-mutations should not occur so often as to corrupt the "normal" behaviour of A-machines.

5. In general, the A-machines almost necessarily form a connected set (in the technical, graph-theoretical, sense) under A-mutation, but this is not important in itself; the important point is that, in principle, proper subsets of the A-set (such as the set of all A-reproducers) may or may *not* be connected under A-mutation. With this understanding, we require that there must exist at least one set of A-machines which is

[1] Note that this does not involve an infeasibly strong notion of "identity" between parent and offspring, but requires only "similarity" to the extent of having all the "same" A-parts in the "same" configuration. These will be formal relationships between formal entities, which can be effectively tested for identity; in itself this says nothing about the capabilities of real, physical, systems. In the terminology of (McMullin 1992a), it can be roughly regarded as a formalisation of the *possibility* of the preservation of S-class in S-descent. Compare also the discussion in (McMullin 1992c, pp. 15–16).

> connected under A-mutation, whose elements are all A-reproducers, and which includes elements having a "wide" (preferably "infinite") range of *A-complexity* (or *A-knowledge*—I shall use the terms interchangeably here). This notion of A-complexity or A-knowledge is necessarily *in*formal, but I shall interpret it roughly as the ability of an A-machine to predict some relevant aspects of the behaviour of its world, and to effectively exploit these predictions in conditioning its interactions with that world (for more detailed discussions, see McMullin 1992b, pp. 5–7, and McMullin 1992e, Chapter 3). The general idea of connectivity under some kind of mutational relationship is closely related to what Kauffman (1990) has called "evolvability"; essentially the same issue has also been previously discussed (in a specifically biological context) by Maynard Smith (1970).

Taken together, these at least approximate to a minimum set of necessary conditions for the growth of automata complexity (if such growth is to occur spontaneously, by Darwinian evolution). More specifically, we must have A-constructors which can at least *maintain* A-complexity (A-reproducers being a special case of this), for S-actors have this property, and only S-actors can give rise to S-lineage selection; and we must have some mechanism, over and above this, corresponding to S-creation, whereby A-complexity may actually *increase* (McMullin 1992a).

This is, of course, precisely the rationale for formulating this particular set of conditions; but I reiterate that, *even* if all these conditions can be satisfied, they are not *sufficient* for the growth of A-complexity. This point will be returned to subsequently. For the moment, we note that, *prima facie*, it is not at all clear that the conditions already identified can be satisfied, even in principle—i.e. that any A-system satisfying these conditions exists. Von Neumann put the issue this way:

> Everyone knows that a machine tool is more complicated than the elements which can be made with it, and that, generally speaking, an automaton A, which can make an automaton B, must contain a complete description of B and also rules on how to behave while effecting the synthesis. So, one gets a very strong impression that complication, or productive potentiality in an organization, is degenerative, that an organization which synthesizes something is necessarily more complicated, of a higher order, than the organization it synthesizes.
>
> von Neumann (1966a, p. 79)

If this were really so it would represent, at the very least, a severe difficulty for the continued application of reductionist, or mechanistic, theories in biology. It is evidently an issue of considerable and profound importance.

So, the question becomes: can we actually exhibit an A-system which demonstrably *does* meet all the conditions stated above?

Von Neumann's crucial insight was to recognise that there *is* a way whereby this can be done (at least in principle), and done relatively easily at that. I shall outline his argument in the following sections; but I must stress, in advance, that von Neumann does *not* claim that the biological world necessarily or exactly conforms to the particular axiomatizations, or architectural organisations, which he describes. That is, von Neumann does not claim that his solution to P_v is, in any sense, *unique*; rather, his demonstration must be regarded only as a proof of the *principle* that a solution is possible at all, and thus as leaving open the possibility of *some* valid, strictly reductionist (A-systematic), theory of the biological world—even if its *detailed* mechanisms are found to be different, perhaps even radically different, from von Neumann's example.

6.2.2 Alan Turing: the A$_T$-system

Von Neumann's attempted solution to P_v was heavily, and explicitly, influenced by Turing's formulation and analysis of a certain formalised class of "computing machines" (Turing 1936). However, the relationship between these analyses of von Neumann and Turing can be easily misunderstood, and will therefore require careful examination.

Turing's analysis had the following general structure. He first introduced a basic formalization of the notion of a *computing* machine. In my terms, this corresponds to the definition of a (more or less) specific A-system. I shall distinguish references to this with a subscript T, thus: A$_T$-system, A$_T$-machine etc. What I term an A$_T$-machine is, of course, what is more commonly referred to as a *Turing Machine* (e.g. Minsky 1967; Lewis & Papadimitriou 1981).

One of Turing's major results was that, in a perfectly definite sense, certain particular A$_T$-machines can be so configured that they can *simulate* the (computational) operations of *any* A$_T$-machine—and can thus, in a definite sense, realise the same "computation" as any A$_T$-machine. Turing called any A$_T$-machine having this property a *universal* (computing) machine. Von Neumann referred to this same property as "logical universality" (von Neumann 1966b, p. 92). It should be clear that this *concept* (though not, of course, any particular automaton) can be generalised across *any* A-system which supports some notion of "computing automaton", in the following way. Call any "computation" which can be carried out by some A-machine an *A-computation*; then, a "universal logical (computational) machine", which I shall term simply a *ULM*, is a single A-machine which, when suitably "configured", can carry out *any* A-computation.

Note carefully that (so far, at least), there is no claim about any relationship which might exist between A-computations (and thus ULMs) in *different* A-systems. The ULM concept is well defined only relative to a particular A-system (and especially the particular notion of A-computation incorporated in that A-system).

We may restate Turing's claim then as a specific claim for the existence of at least one ULM within the A_T-system—i.e. the existence of a ULM_T. Again, what I call a ULM_T is now most commonly referred to as a *Universal Turing Machine* (Minsky 1967; Lewis & Papadimitriou 1981). An essential concept in Turing's formulation of his ULM_T is that its operations are "programmed" by a list of "instructions" and that, as long as a fairly small basis set of instructions are supported, it is possible to completely describe the computational behaviour of an arbitrary A_T-machine in terms of a finite sequence of such instructions. That is, a ULM_T is made to simulate the computations of any arbitrary A_T-machine simply by providing it with an appropriately coded *description* of that machine.

Note that, in itself, Turing's claim for the existence of at least one ULM_T is entirely neutral as to whether ULM's can or do exist in any other A-system, or, more generally, whether "computing machines" in general share any interesting properties across different A-systems. These are important issues, which were central to the problem which Turing was attempting to solve. However, although von Neumann was, in some sense, inspired by Turing's work on the A_T-system, his *problem* was entirely different from Turing's problem; and, as a result, I claim that these issues were more or less irrelevant to von Neumann's work.

6.2.3 On "Universal" Construction

Turing formulated the A_T-machines specifically as *computing* machines; the things which they can manipulate or operate upon are not at all the same kinds of things as they are made of. No A_T-machine can meaningfully be said to *construct* other A_T-machine(s)—there are no such things as A_T-constructors or, more particularly, A_T-reproducers.

Von Neumann's basic idea was to generalise Turing's analysis by considering abstract machines which *could* operate on, or manipulate, things of the "same sort" as those of which they are themselves constructed. He saw that, by generalising Turing's analysis in this way, it would be possible to solve P_v in a very definite, and rather elegant, way.

In fact, von Neumann considered a number of distinct A-systems, which are not "equivalent" in any general way, and which were not always completely formalised in any case. However, a key thread running throughout all this work was to introduce something roughly analogous to the general concept of a ULM, but defined relative to some notion of "construction" rather than "computation".

Von Neumann's new concept refers to a particular kind of A-machine which he called a *universal constructor*; I shall refer to this as a "universal constructing machine", or *UCM*.

The analogy between the ULM and UCM concepts is precisely as follows. Like a ULM, the behaviour of a UCM can be "programmed", in a rather general way, via a list of "instructions". In particular, these instructions may provide, in a suitably encoded form, a *description* of some A-machine; and in that case, the effect of "programming" the UCM with that description will be to cause it to *construct* the described A-machine (assuming some suitable "environmental" conditions: I shall have more to say about this requirement later).

Thus, just as a ULM can "simulate the computation of" *any* A-machine (when once furnished with a description of it), so a UCM should be able to "construct" *any* A-machine (again, when once furnished with a description of it, and, of course, always working within a particular axiomatization of "A-machine", which is to say a particular A-system).

We may trivially note that since there do not exist any A_T-constructors at all, there certainly does not exist a UCM_T, i.e. a UCM within the A_T-system.

I emphasise strongly my view that it was precisely, and solely, the *spanning of all A-machines in a particular A-system* that mandated Turing's original usage of the word "universal" (in "universal machine", or ULM_T in my terms), and which also mandated von Neumann's analogous usage (in "universal constructor", or UCM in my terms). The characteristic operations of the two kinds of machine (computation and construction, respectively) are quite different.

In Turing's original paper (Turing 1936) he argued, *inter alia*, that there exists a ULM_T, in the sense already described—a single A_T-machine which can simulate (the computations of) any A_T-machine. This is a technical, formal, result—a *theorem* in short—which Turing *proved* by actually exhibiting an example of a specific A_T-machine having this property. We shall see that von Neumann sought to achieve an essentially analogous, perfectly formal, result for a UCM—i.e. to prove the existence of such things, at least within some "reasonable" A-system, and to do so by precisely paralleling Turing's procedure, which is to say by actually exhibiting one. At this level, the analogy between these two developments is very strong and direct, and the word "universal" has a clearly related implication in both "UCM" and "ULM" within their respective domains.

However, a problem arises because the "universal" in "ULM" actually admits of three (or perhaps even five, depending how they are counted!) quite distinctive interpretations or connotations—only *one* of which is the one described above as being legitimately preserved in von Neumann's intended analogy. If one mistakenly supposes that any of the *other* connotations should be preserved (as well as, or instead of, the correct one) then the result can be serious confusion, if not outright error.

I have provided a properly detailed and exhaustive account of this issue elsewhere (McMullin 1992e, Section 4.2.4). In particular, I attempt to detail the negative influence of this confusion, after von Neumann himself put the work aside. But to repeat that detailed discussion here would take me too far afield. For my present purposes, it is sufficient to summarise my claims as follows:

- Von Neumann introduced the notion of a UCM, by analogy with Turing's ULM_T, as a particular kind of A-machine which could, when suitably programmed, construct *any* A-machine.
- This notion *only* becomes precise in the context of a *particular* axiomatization of A-machines, i.e. a particular A-system (and A-set).
- The UCM concept, as originally formulated by von Neumann, does not *inherently* involve any comment about the "computational" powers either of itself or of its offspring, and does not involve or imply any "natural" generalisation of the Church-Turing Thesis.

6.2.4 von Neumann's Solution

6.2.4.1 The Kinematic Model

> A complete discussion of automata can be obtained only by ... considering automata which can have outputs something like themselves. Now, one has to be careful what one means by this. There is no question of producing matter out of nothing. Rather, one imagines automata which can modify objects similar to themselves, or effect syntheses by picking up parts and putting them together, or take synthesized entities apart. In order to discuss these things, one has to imagine a formal set-up like this. Draw up a list of unambiguously defined elementary parts. Imagine that there is a practically unlimited supply of these parts floating around in a large container. One can then imagine an automaton functioning in the following manner: It also is floating around in this medium; its essential activity is to pick up parts and put them together, or, if aggregates of parts are found, to take them apart.
>
> von Neumann (1966a, p. 75)

Von Neumann's initial, informal, attempted solution to P_v was first presented in a series of lectures given to a small audience at the Princeton Institute for Advanced Studies, in June 1948; no formal record of these lectures survives, but Burks reconstructed much of the detailed exposition from notes and memories of the audience (Burks 1966b, p. 81). Von Neumann himself recounted the ideas, though in somewhat less detail, at the Hixon symposium in September 1948 (von Neumann 1951), and during his lectures at the University of Illinois in December of the following year (von Neumann 1966a). These presentations were all based on what came to be called his *kinematic* model.

This model involved something of the order of 8–15 distinct, primitive, A-parts, visualised as mechanical components freely floating in a two or three dimensional Euclidean space. These included basic structural elements ("rigid members" or "girders"), effectors ("muscles", "fusing" and "cutting" organs), and elements to realise general purpose signal processing ("stimulus", "coincidence", and "inhibitory" organs). Sensors could be indirectly realised by certain configurations of the signal processing elements. Roughly speaking, any more or less arbitrary, finite, aggregation of these primitive parts, mechanically attached to each other, would then qualify as an A-machine in this system.

In this basic model von Neumann intended to disregard all the detailed problems of mechanics proper—force, acceleration, energy etc.—and restrict attention to essentially geometrical-kinematic questions; which is why Burks introduced the term *kinematic* to identify this kind of model (Burks 1966b, p. 82).

The kinematic model was never formalised in detail; indeed, to do so would involve overcoming quite formidable obstacles. However, even in a very informal presentation, the model does provide an intuitive picture supporting the arguments von Neumann wished to present. I shall more or less follow von Neumann in this. Thus, the following discussion of von Neumann's solution to P_v is actually phrased in completely abstract terms, with no explicit reliance on the kinematic (or any other) model; but it may nonetheless help the reader's intuitive understanding to imagine, in the first place at least, that its terms are interpreted relative to the kinematic model.

Also following von Neumann (though perhaps rather more so than he), I adopt a certain amount of mathematical, or quasi-mathematical, notation here. This should not be taken too seriously; it is essentially a shorthand device, intended only to render certain elements of the argument as clearly and concisely as possible. There is no question that I provide anything which could be regarded as a *proof*, in a formal, mathematical, sense—the notation notwithstanding.

6.2.4.2 Some Notation

Denote the ("universal") set of all A-machines in some particular A-system by M.

In general, the "combination" or "composition" of A-machines (primitive A-parts, or otherwise) will be denoted by the symbol $\oplus$. That is, if m_1 and m_2 are two A-machines, then $(m_1 \oplus m_2)$ will denote a single A-machine consisting of m_1 and m_2 "attached" to each other. For the purposes of this outline, it will be assumed that such compositions are always well-defined, in the sense that, for arbitrary $m_1, m_2 \in M$, there will exist some unique $m_3 \in M$ such that $(m_1 \oplus m_2) = m_3$. The precise nature or mechanism of such "attachments" might, in general, be ambiguous; but I shall assume that that extra complication can be overcome in any particular A-system.

Constructional processes in the A-system will be denoted by the symbol $\rightsquigarrow$; that is, if an A-machine m_1 constructs another A-machine m_2, separate from itself, then this will be written $m_1 \rightsquigarrow m_2$. Thus, in particular, if some $m \in M$ is an A-reproducer, it must be the case that, under "suitable" circumstances, $m \rightsquigarrow m$.

We require that the A-system should support the existence of a certain special class of A-machine, which can function as "data storage" devices. These will be termed *A-tapes*. The set of all A-tapes will be denoted T. T will, of course, be a proper subset of M. It is an essential, if implicit, property of A-tapes that they are, in some sense, *static*; an A-tape may potentially be transformed into another, different, A-tape (or, if one prefers, the "content" of a "single" A-tape may be altered to a different "value"), but *only* through the action of some other, attached, A-machine (which is not, in turn, an A-tape).

Suppose that a particular UCM, denoted u_0, can be exhibited in this A-system (i.e. $u_0 \in M$), where "programming" of u_0 consists in the composition of u_0 with some A-tape. The A-tape is thus interpreted as encoding a formal description of some A-machine, in some suitable manner ("understood" by u_0). Any A-tape which validly encodes a description of some A-machine (relative to u_0) will be called an *A-descriptor*. We require (from our assertion that u_0 *is* a UCM) that $\forall\ m \in\ M$ there must exist at least one element of T which validly describes m. Thus we can define a function, denoted $d()$ (read: "the A-descriptor of") as follows:

$$\begin{array}{lllll} d & : & M & \rightarrow & T \\ & & m & \mapsto & d(m) \quad \text{s.t.} \quad (u_0 \oplus d(m)) \rightsquigarrow m \end{array}$$

That is, u_0 composed with (any) $d(m)$ will construct (an instance of) m.

We assume that the behaviour of u_0 is such that, when any $(u_0 \oplus d(m))$ completes its constructional process, it will be essentially unchanged (will revert to its original "state"); which is to say that it will then proceed to construct another instance of m, and so on.[2]

The set of A-descriptors is clearly a subset of the set of A-tapes, T; it may, or may not, be a *proper* subset.[3] In fact, we do *not* (for the moment) require any one-to-one correspondence (for example) between the A-descriptors and A-tapes; which is to

[2] I note, in passing that, on the contrary, von Neumann *originally* assumed that the attached A-descriptor would be "consumed" or destroyed when processed by a UCM. However, it turns out that this has no essential significance; it also complicates the subsequent development, and obscures the biological interpretation of von Neumann's ideas. Indeed, von Neumann himself subsequently adopted (in his cellular model) the convention I have adopted here from the first.

[3] That is, it is not clear whether, in the definition given of $d()$, T should be technically regarded as its *range*, or merely a sufficiently inclusive *target*.

say that while every A-descriptor will be an A-tape, the converse will not necessarily hold. In particular, some A-tapes may not validly describe *any* A-machine. The composition of such an A-tape with u_0 is still well-defined (i.e. is some particular A-machine) of course, but we say nothing in particular about the *behaviour* of such a composition.

6.2.4.3 The Core Argument

The UCM u_0 is, of course, introduced as a tool for the solution of P_v; but, to anticipate somewhat, it will turn out that u_0 does *not* (directly) solve P_v. Instead, we shall see that the existence of u_0 "almost" solves it, or, at least, it solves certain aspects of it. Nonetheless, this "near" solution is the very heart of von Neumann's argument. Its deficiencies are relatively minor and can, as von Neumann demonstrated, be relatively easily corrected; but these corrections will make no sense at all until the basic underlying argument—the "near" solution of P_v—is clearly understood. It is the underlying argument that will be elaborated in this section.

Recall that, by definition, u_0 can construct *any* A-machine; therefore, it can construct (an instance of) u_0 itself, when once provided with the relevant A-descriptor, namely $d(u_0)$. Thus, it seems that any UCM should more or less directly yield an A-reproducer, simply by programming it with its own description. I hasten to add that the logic here is actually mistaken, and it is as a consequence of this that u_0 will only "almost" solve P_v; but we shall ignore this for the time being.

Now this result (that u_0 directly implies the existence of a particular A-reproducer) is, *in itself,* almost entirely without interest: for the point is not to exhibit self-reproduction as such, but rather to exhibit the possibility of a spontaneous growth in A-complexity (by Darwinian means). The existence of at least one design for an A-reproducer is certainly a necessary precondition for any solution of this problem; but what we *really* need is the existence of a *set* of distinct A-reproducers, spanning a diverse (preferably "infinite") range of A-complexity; which set must also be connected under some reasonable definition of A-mutation. u_0 *on its own* does not yield this.[4]

However, it turns out (and this is one of von Neumann's crucial insights) that the argument for u_0 giving rise to a single A-reproducer could (if it were valid) be immediately extended, in the following manner.

Let X be the set of all A-machines having the property that any $x \in X$ can be composed with u_0 without "interfering" with the basic operation of the latter. That is, given any A-machine of the form $(u_0 \oplus x)$, it will still be possible to compose this with any A-descriptor and the effect will be that the composite A-machine will still be able to construct the described A-machine; more concisely, we assume, or require, X to be such that:

$$\begin{aligned}&\forall\, m \in M,\\ &\forall\, x \in X,\\ &\qquad\qquad ((u_0 \oplus x) \oplus d(m)) \rightsquigarrow m\end{aligned}$$

Any composite A-machine $(u_0 \oplus x)$ may, of course, be capable of doing other things as well. In particular, we assume that it can do essentially any of the things which the "isolated" A-machine x was able to do. This is a roundabout way of saying that we assume that the A-complexity of any composite A-machine of the form $(u_0 \oplus x)$ is at least as great as either u_0 or x taken separately (whichever of the latter two A-complexities is the greater).

We make one further, critical, assumption about the set X: we require that it include elements spanning a "wide" (preferably "infinite") range of A-complexity. This is, strictly, a new and independent assumption. However, we may hope that it will not be *too* difficult to satisfy, assuming that the set M satisfied such a condition in the first place—which presumably it will, provided we choose our axiomatisation "reasonably". That is, while we do not expect to have $X = M$ as such, we can reasonably suppose that if M itself offers a very large set of A-machines having a very wide variety of behaviours (A-complexity) then there should "surely" be a subset, still spanning a wide variety of behaviours, but whose elements do not interfere with the behaviour of u_0.

Now, by hypothesis, every A-machine of the form $(u_0 \oplus x)$ can still, by being suitably programmed, construct any arbitrary A-machine. That is to say, we have gone from having a *single* UCM u_0, to having a whole family or set of "related" UCMs ("related" in the sense of having the same "basic" UCM, u_0, embedded within them—which means, *inter alia*, that they all process the same description language, or are all compatible with the same set of A-descriptors). I shall denote this set of related UCMs by U:

[4] To put the same point conversely: if we were merely interested in self-reproduction "as a problem in itself" (of course, we are not!) then any A-reproducer at all would do, and the introduction of u_0 would be unmotivated, if not positively counterproductive; it is plausible (I might even say *likely*) that there are far easier ways to design a single A-reproducer than by trying to base it on anything as sophisticated as a UCM!

$$U = \{(u_0 \oplus x) | x \in X\}$$

As a special case I stipulate that u_0 itself is also a member of U.

Now the elements of U are *not* themselves A-reproducers; but since every element *is* a UCM in its own right then, if the original argument applied to u_0 were valid (and we shall return to *this* issue shortly), every element of U implies or gives rise to a distinct A-reproducer merely by programming it with its own description.

Thus, corresponding to every $x \in X$ there exists a (putative) A-reproducer which effectively contains x as a (functional) subsystem (and is therefore, presumably, to be considered at least as A-complex as x). Which is to imply that the existence of u_0 does not merely yield a single (putative) A-reproducer; instead, with the addition of some more or less innocuous additional assumptions (i.e. those relating to the existence and properties of the A-machines making up the set X) u_0 implies the existence of a whole set of A-reproducers, spanning the requisite range of A-complexities.

With this observation we are now very close to a solution of P_v. But a question still remains as to the relationships between these A-reproducers under A-mutation: that is, have we any basis for claiming that this set of A-reproducers, anchored on u_0, will be connected under any plausible interpretation of A-mutation?

Well, note that any of these A-reproducers can be effectively transformed into any other simply by appropriate change(s) to the A-tape. In more detail, if we regard A-mutation as including the possibility of a spontaneous change in the A-tape, changing it from being an A-descriptor of any one A-reproducer (based on some $u_1 \in U$) to being an A-descriptor of some other A-reproducer (based on some $u_2 \in U$), then the future offspring of the affected A-reproducer will incorporate (instances of) u_2 instead of u_1, and will then reproduce as such. As a general principle, it would seem that any A-mutation to the A-tape which did not affect the construction of the embedded (instance of) u_0 in the offspring (i.e. any A-mutation not affecting the $d(u_0)$ "section" of the A-descriptor) would be at least a candidate for this. So it seems at least "plausible", that the set of A-reproducers, anchored on u_0, might indeed be *connected* under some relatively simple notion of A-mutation applied to the A-tapes.

Strictly, it must be carefully recognised that this last claim does involve *some* assumption about the encoding of A-machine descriptions which is "understood" by the particular UCM, u_0 (and thus by all the UCMs in U). So far, I have said that, for every A-machine, there exists at least one A-descriptor which describes it (relative to u_0); but I have not said how "dense" this set of A-descriptors is within the set of all A-tapes; nor, more particularly, have I said how dense is the *subset* of A-descriptors which validly describe the elements of the set of A-reproducers anchored on u_0. Specifically, one can imagine encodings which would be very "sparse"—i.e. such that "most" A-tapes are not A-descriptors of any such A-reproducer, and, therefore, such that an A-mutation of an A-descriptor, defined as affecting only a single A-part, would be unlikely to yield an A-descriptor of any other A-reproducer, but would rather yield some kind of more or less "nonsensical" A-tape. However, one can equally imagine encodings which *are* dense in this same sense. For the time being at least, we are thus free to *assume*, or stipulate, that the encoding in use is of just this sort. Like all our other assumptions (pre-eminently the existence of u_0 itself) this can ultimately be defended *only* by showing that it can be satisfied in some particular A-system.

At this point then we have, based essentially on the assumed existence of a UCM u_0, a tentative schema for the solution of P_v. It must be emphasised that this schema depends critically on the construction universality of u_0. It would not, for example, be possible to formulate a similar schema based on any arbitrary A-reproducer, of unspecified internal structure—for such an arbitrary A-reproducer could not generalise to a *set* of A-reproducers of essentially unlimited (within the scope of the A-system itself) A-complexities; nor could such an arbitrary A-reproducer offer any systematic form of A-mutation which could be expected to connect it with other A-reproducers.[5]

It is thus clear, once again, that the problem P_v is utterly different from the (pseudo-)problem of self-reproduction "in itself"; for whereas the UCM con-

[5] This is perhaps a more subtle point than can be properly done justice to here. The critical thing is that by thinking of A-mutation as occurring in the space of *A-descriptors*—which involves an essentially *arbitrary* encoding of the A-machines—we can quite reasonably require that the encoding be *designed* to be just such that the images (A-descriptors) of our putative A-reproducers should be as close as we like to each other in this space, thus (indirectly, via construction) yielding the necessary A-mutational connectivity of the A-reproducers themselves. But no such assumption of connectivity could be justified if we think of the A-mutations as affecting some essentially arbitrary set of A-reproducers *in general*, for we then have no basis for supposing they are, or can be made to be, "close" to each other in any relevant space. See the further discussion of this point in section 6.3.2 below.

cept is seen (for the time being at least) as central to the solution of P_v, its introduction would be gratuitous, if not unintelligible, if one thought the problem at hand were merely that of self-reproduction.

This completes the presentation of von Neumann's core argument; we must now turn to criticism and elaboration of it.

6.2.4.4 A Minor Blemish(?)

I pause to identify and correct a logical error in the core argument thus far presented. I should emphasise that von Neumann himself presented his theory only in its final, corrected, form. I have chosen to present it first in a (slightly) mistaken form because I think this can help to clarify the relative importance and significance of the various elements of the argument.

I refer to the error merely as a "minor blemish" because an essentially minor modification of the argument can correct it; but I do not mean by this to imply that it was "easy" to correct *in the first instance*. Even though the required modification ultimately proves to be minor, it arguably required a remarkable insight on von Neumann's part to see that a correction was possible at all, never mind actually formulating such a correction. I admit all this. But I want to emphasise that, in my view, von Neumann's *central* achievement is already contained in what I have called the core argument—compared to which the technical correction introduced in this section, though strictly necessary of course, is a very minor matter indeed. I point this out because at least some commentators seem to have supposed, on the contrary, that the mere "trick" to be introduced here was of the essence of von Neumann's analysis—see, especially, Langton's discussion (Langton 1984, pp. 136–137), and, to a lesser extent, Arbib (1969, pp. 350–351).

The logical error is this: in the original development, it was stated, or assumed, that, given an arbitrary UCM u, then there will exist a corresponding A-reproducer, consisting simply of u programmed with its own A-descriptor—i.e. the A-machine $(u \oplus d(u))$. This is simply false.

What we actually have here is:

$$(u \oplus d(u)) \leadsto u$$

whereas, what we would strictly require for self-reproduction would be something like:

$$(u \oplus d(u)) \leadsto (u \oplus d(u))$$

which is clearly not the case.

In words, the A-machine $(u \oplus d(u))$ constructs, not another instance of itself, but an instance of the "naked" A-machine u, with no A-tape attached. This is clearly not self-reproduction. This flaw applies, of course, to u_0 itself, but equally to all the other elements of the set U. *None* of them imply the existence of an A-reproducer, in the manner indicated; which is to say that none of the original, putative, A-reproducers are actually self-reproducing, and the proposed schema for solving P_v fails utterly.

We are now ready to consider von Neumann's mechanism for getting around these difficulties. Von Neumann presented this (within the kinetic model) essentially in terms of a modification of the UCM u_0, while leaving the formal description language more or less unchanged. For reasons which should become quickly apparent, I shall refer to this new modified kind of A-machine as a "Universal Genetic Machine" or *UGM*, though these are not terms which von Neumann himself ever used. I note that the UGM is (or, at least, can be) defined not as something *different* from a UCM, but as a special *kind* of UCM—a UCM subject to a certain constraint, to be explained below, on the description language which it supports. This roughly underlies Burks' (1966a, pp. 294–295) development (or "completion") of von Neumann's ideas and explains why both Burks (1970b, p. xi) and Arbib (1969, Chapter 10), for example, can use the term "universal constructor" synonymously for the two kinds of A-machine I distinguish as UCMs and UGMs.

Although von Neumann originally introduced the UGM as, literally, a modification of a UCM, nothing crucial hangs on this procedure. That is, it may, or may not, be the case, in a particular A-system, that if a UCM exists at all, it can be "easily" modified to yield a UGM. So, technically, rather than relying on any such implication, I now simply *strengthen* the original requirement that our A-system support "some" UCM, and demand *instead* that it specifically support a UGM as such. So: we suppose that our UCM u_0, of the previous sections, is now constrained to be, in fact, a UGM.

Since u_0 is still a UCM we know that, given any A-machine $m \in M$, there must exist an A-descriptor $d(m)$ which would cause u_0 to construct (an instance of) m. However, we will make at most informal or heuristic use of this property. The important property of u_0 is the constraint on its description language which is introduced by virtue of its being a U*GM*, and this is as follows. Given any A-machine $m \in M$, there must exist some A-descriptor $d'(m)$ which would cause u_0 to construct (an instance of) $(m \oplus d'(m))$. More formally, we are declaring the existence of a function, denoted $d'()$ (read: "the dashed A-descriptor of") with the following definition:

$$\begin{array}{llll} d' & : & M & \rightarrow & T \\ & & m & \mapsto & d'(m) \ \text{s.t.} \\ & & & & (u_0 \oplus d'(m)) \leadsto (m \oplus d'(m)) \end{array}$$

Before showing how this property can resolve the difficulty with achieving self-reproduction, we need to provide some argument to suggest that such a property *might* actually be realisable. Informally, the idea is that each $d'(m)$ can contain, embedded within it, the A-descriptor $d(m)$; faced with $d'(m)$, u_0 first identifies this embedded A-descriptor $d(m)$ and decodes it, "as usual", to construct the described A-machine; but u_0 then goes on to construct a *copy* of the complete A-descriptor $d'(m)$, and attach it to the offspring A-machine m. The $d'(m)$ A-descriptors can thus simply be the original $d(m)$ descriptors with some kind of qualifier or flag added to indicate that this extra copying step should be carried out.

Another way of looking at this is that u_0 now, as it were, supports two different formal languages: the original one (which can still be freely designed to satisfy any particular requirements we like—such as ensuring that the A-descriptors of certain A-machines will be A-mutationally "close"to each other); and a new, impoverished language, which can code *only* for A-tapes, and which uses the simple coding that every A-tape is its own A-descriptor. By *alternately* interpreting an attached A-tape in these two *different* ways (whenever the A-tape is flagged to indicate that this is desired), u_0 can ensure that, for every $m \in M$ there will correspond an A-descriptor, $d'(m)$, describing precisely the composite A-machine $(m \oplus d'(m))$.

Now, given this property of u_0, we *can* directly identify a corresponding A-reproducer—*not* by programming it with its A-descriptor $d(u_0)$, but by programming it with its *dashed* A-descriptor $d'(u_0)$. By definition, this is the A-descriptor of $u_0 \oplus d'(u_0)$. That is:

$$(u_0 \oplus d'(u_0)) \leadsto (u_0 \oplus d'(u_0))$$

and, at last, we have genuine self-reproduction.

The rest of the core argument can now be completely rehabilitated; assuming that all the A-machines $x \in X$ still have the property of not interfering with the basic operation of u_0 (when composed with it) we can say that all the machines $u \in U$ will be, not merely UCMs, but UGMs. Just as with u_0 then, each $u \in U$ will give rise to a corresponding A-reproducer by programming it with the A-descriptor $d'(u)$. The complete core argument can then go through, yielding a now valid solution schema for P_v.

6.2.4.5 Loose Ends(?)

I have deliberately termed what has so far been achieved a solution *schema* for P_v, rather than a solution proper. It suggests, in outline, a method whereby we might establish that an A-system satisfies the requirements set out in the statement of P_v: but it does not, in itself, identify any particular such A-system. There are, that is to say, some decidedly loose ends to be tidied up before P_v can properly be declared solved.

Nonetheless, before proceeding to these loose ends, I wish to make clear that, in my view, this is a relatively routine or minor task. It seems to me that the core argument (as it has now been presented) satisfactorily solves all the *substantive* difficulties bound up with P_v; tidying up loose ends is a necessary drudgery of course, but further, real, progress cannot now be expected before we can carry out a critical reformulation of our problem situation (in the light of having *solved* P_v).

The loose ends in question here amount essentially to the exhibition of a particular A-system which meets the requirements for the core argument to be applied to it. Von Neumann perhaps hoped originally to develop the kinematic model to a point where this would be possible. Be that as it may, he instead turned his attention to what Burks (1966b, p. 94) calls his *cellular* model—a form of cellular automaton.

The questions to be answered for this particular A-system may be conveniently divided into one which is purely formal, and a second which is largely informal:

1. The formal question is whether there exists a basic UGM u_0, and a set of related UGM's U, such that the (dashed) A-descriptors associated with the corresponding A-reproducers are "dense" (in the sense of being connected under A-mutation) in the space of A-tapes. Once the particular A-system is properly formalised, these things become matters of fact, accessible (in principle at least) to formal proof. The attempt to provide such proofs constituted the larger part of von Neumann's unfinished manuscript *The Theory of Automata: Construction, Reproduction, Homogeneity* (von Neumann 1966b).

2. The informal question is whether the identified A-reproducers span the requisite range of A-

> complexity. Since A-complexity itself is an informal concept here, any answer to this will necessarily be informal. Von Neumann himself did not attempt to explicitly answer this question for his cellular (or, indeed, any other) model; perhaps he would have done so in completing his manuscript; or perhaps he considered that an affirmative answer was self evident. In any case, I shall give a brief discussion of this issue, because it is in my view an important, albeit somewhat intractable, question, and it seems that this has not generally been appreciated.

There are, of course, many other questions which could be taken up in a completely comprehensive account. For example, we should perhaps discuss critically whether von Neumann's cellular model *does* provide a "reasonable" axiomatization of the notion of "automaton" at all;[6] or at least we should consider whether the model satisfies the requirements of not having "too many" primitive A-parts, which are not individually "too complex" etc. But these issues would take me too far afield, and I shall therefore restrict myself here to the two questions explicitly identified above, which I consider to be most immediately relevant to the topics at hand.

The first question relates to the design of a basic UGM, and the development of this to establish a diverse set of A-reproducers, which is connected under A-mutation of the A-descriptors.

The first part of this question—the design of the basic UGM—has been addressed positively several times over. Von Neumann himself had more or less completed the demonstration that a basic, minimal (i.e. with no additional functionality) UGM exists in his cellular model (by exhibiting the design for a particular u_0) at the time he put his manuscript aside. Burks (1966a) showed in detail how this demonstration could be completed, and also outlined how the design could be significantly simplified. Thatcher (1970) has demonstrated a detailed version of this simplified design. Codd (1968) has exhibited a basic UGM design in a different cellular model, having only 8 states per cell (compared to the original 29 states per cell in von Neumann's model); and Berlekamp *et al.* (1982) have argued, without detailing a design, that a UGM is possible in a particular cellular model having only 2 states per cell (Conway's so-called "Game of Life"). Although all of these represent arguments "in principle"—no fully fledged UGM-based A-reproducer has actually been built or demonstrated, to my knowledge—the arguments are, overall, satisfactory and we can take it that the possibility of exhibiting a basic UGM (and thus a basic A-reproducer) within a suitably "simple" (cellular) model (von Neumann's or otherwise) is now well established.

The remaining parts of the first question—identifying the set X of A-machines which could be composed with the given u_0 without compromising its operations, and of establishing the connectivity of the corresponding A-reproducers under A-mutation—have, on the other hand, received little or no explicit attention. Von Neumann himself seemed loosely to talk in terms of X being essentially coextensive with M—i.e. neglecting the possibility that there would be any interference with the operation of u_0 (von Neumann 1966b, pp. 119, 130–131); similarly he did not seem to give any explicit argument to support the A-mutational connectivity of the A-reproducers. Subsequent commentators do not seem to have added anything further. My disagreement with leaving matters in this state is minor, though not quite pedantic.

Firstly, for the sake of precision or completeness I think it should be explicitly recognised or admitted that X will (almost certainly) *not* be coextensive with M. But, equally, I do not think it generally feasible to give any better characterisation of X than simply to say that the elements of U are indeed still UGMs in their own right (i.e. my definition of X is purely existential—it offers no clue as to how, for example, one might systematically *generate* the elements of X other than by simply *testing* elements of M in turn). In the case of von Neumann's cellular model (or, indeed, his kinematic model) I am quite willing to accept, without any attempt at proof, that although X cannot be coextensive with M, it is still an infinite set, spanning *essentially* the same range of A-complexity as M itself—and *this* is really the critical point. It is, perhaps, so obvious that von Neumann simply felt it was not necessary to say it. As to whether the range of A-complexity offered by M in the first place is, informally, sufficient for a solution of P_v, that relates to question 2 above, and I shall take it up separately, in due course.

The second outstanding aspect of question 1 follows on from the status of X: we wish to establish that the set of A-reproducers anchored on U (which is to say, indirectly anchored on X) is connected under some specified interpretation of A-mutation (of the A-descriptors). A formal answer to this might, in principle, be possible; but would be exceedingly difficult, and has never, to my knowledge, been at-

[6] Thus, for example, Kampis & Csányi (1987) argue that the self-reproduction phenomena (SR) at least, exhibited by von Neumann, "cannot avoid a sort of triviality and in this they are basically different from real SR, such as that of living organisms".

tempted. It would require *inter alia* that we be able to characterise the set X much more precisely that heretofore—a task which I have just accepted as being very difficult, if not impossible, in itself.

I think the best we can reasonably do (and this is actually very good, albeit far short of a formal proof) is the following:

- We can require that the formal description language supported by u_0 incorporate some degree of "compositionality"; specifically, we require that the "portion" of the A-descriptors coding for the "core" part of the A-reproducers (i.e. coding for the u_0 subsystem itself) can be, to a greater or lesser extent, "separated" out. I mean by this that there will exist many possible A-mutations (namely any affecting any *other* portion of the A-descriptors, and thus affecting only the x subsystem of the offspring) which would not compromise this essential core of the offspring. This greatly enhances the possibility that such an A-mutation will, indeed, yield another A-reproducer, and may be said to have already been implicit in our earlier discussion of the very possibility that the A-reproducers, anchored on u_0, might be connected under A-mutation.

- Furthermore, we can require the language to be such that the portions of the A-descriptors encoding the x subsystem of the offspring should be "dense" *at least relative to* M. That is, while it is difficult, if not impossible, to *directly* guarantee that the encoding will be such that most (or even any) A-mutations of this portion of an A-descriptor will yield an encoding of another $x \in X$ (which is to say, the A-descriptor of another A-reproducer, or the dashed A-descriptor of another $u \in U$), it is perfectly feasible to ensure that most (if not all) such A-mutations at least yield another $m \in M$ (as opposed to simply yielding nonsense—an A-tape not validly describing any A-machine at all). We can now couple this with our earlier (entirely informal) acceptance that, although X cannot be coextensive with M, it will be very large and diverse, to conclude that, even though not all such A-mutations will yield a viable offspring (another A-reproducer) a significant "fraction" plausibly should; and *this* is enough to persuade me (at least) that while the entire set of A-reproducers anchored on u_0 *may* not be connected under A-mutation, some infinite, and diverse, subset of it *will* be; that being the case, I suggest that the requirement involved in solving P_v (namely, that this connected subset span a sufficient range of A-complexities) can still be taken as met (always assuming that M itself spans such a range in the first place).

I should add, of course, that Von Neumann did indeed ensure that the encoding(s) he used were just such that these two conditions are satisfied (see, in particular, von Neumann 1966b, pp. 130–131).

I now come to the last outstanding loose end, my question 2 above. Given the discussion of question 1, question 2 has now resolved itself into the question of the range of A-complexity spanned by the entire "universal" set of A-machines (M) in, say, von Neumann's cellular model; for it has been argued that the (A-mutationally connected) set of A-reproducers, anchored on u_0, will span essentially this same range.

Despite my calling this a mere "loose end", I consider that it is, in its way, quite the hardest question directly associated with the solution of P_v; and since I will not pretend to be able to offer a satisfactory answer, I can be mercifully brief! I think that one somewhat promising approach to this question might be based on relating the informal idea of A-complexity to what Burks (1960) has called the *behavior* of an automaton or A-machine. However, I have reviewed this idea previously (McMullin 1992e, Section 4.2.5.5), and a detailed consideration of it here would take me too far afield. Instead I shall rely on the only answer which I think von Neumann himself could be said to have offered to this question. This is simply to say yes, M does span a sufficient range of A-complexity, *and this is self-evident.* This answer has, at least, the merit of an overwhelming simplicity, and is probably sufficient for my immediate purposes.

6.3 A New Problem Situation

6.3.1 P_a: The Problem of *Autonomy*

6.3.1.1 An Initial Formulation

Von Neumann's formulation and solution of *some* of the fundamental problems underlying the (Darwinian) growth of complexity in artificial systems was a very substantial achievement. But it still falls far short of a *complete* solution of the problems I subsume under the phrase *Artificial Darwinism.* I should therefore like to summarise here my view of the new problem situation which arises as a result of von Neumann's work, and identify, albeit rather crudely, one particular new problem, which I shall call the problem of *autonomy*, or P_a.

Von Neumann (and various successors) established that a UGM could be embedded in his 29-state cellular A-system and, indeed, that the existence of a set of A-reproducers could thus be established which would be connected under A-mutation (albeit no A-mutational *mechanism* was explicitly built into the A-system), and which could fairly reasonably be described as spanning an indefinitely large range of A-complexity. This A-system therefore satisfies *some* conditions which are arguably necessary for the spontaneous growth of A-complexity by Darwinian evolution (which is not, of course, to say that von Neumann's *particular means* of meeting these conditions are "necessary"). Exhibiting this possibility exhausts the scope of P_v, as I defined it.

In this new situation one new question or problem which immediately presents itself is this: will von Neumann's A-system *in fact* exhibit a spontaneous growth in the A-complexity of A-reproducers, by Darwinian evolution (when once "seeded" with an initial A-reproducer)? Indeed, will it exhibit Darwinian evolution of the A-reproducers at all (with or without a growth of A-complexity)?

As far as I am aware, this has never been empirically tested, but there seems to be little doubt as to the outcome which can be expected from such tests: unless special *ad hoc* measures are taken to preempt any substantive interactions between the A-reproducer(s) they will destroy each other quite quickly, and any initial population will become extinct. The population might be sustained, or might even grow, if interactions are effectively prevented, but that would defeat the purpose by preempting natural selection,[7] and thus Darwinian evolution. In any case, there will *not* be any significant Darwinian growth in A-complexity.

It would be mildly interesting to see these predictions tested; but there is good reason for believing that such tests are unnecessary. It seems to be quite clear that all these A-reproducers, in the various (cellular) A-systems I have mentioned thus far, are extremely *fragile*. The self-reproducing behaviour *relies* on the surrounding space being essentially quiescent, and on there being no interference from other, active, configurations. While simple procedures could be adopted such that, from an initial seed A-reproducer, the offspring are all carefully located so as not to interfere with each other, or their subsequent offspring etc., this would preempt the kind of direct and indirect interactions which are essential to the operation of natural selection. If, on the contrary, more or less unrestrained interactions were allowed, the A-reproducers would very quickly destroy each other, and make the environment uninhabitable. The basic von Neumann design of genetic A-reproducer, and comparable designs for the other cellular A-systems, whatever their positive merits (and they are substantial, as we have seen), lack any capability to protect or maintain their own integrity in the face of even minor perturbations. In my view therefore, they could not possibly survive in any but the most strictly controlled environments; which is to say that they could not effectively demonstrate the operation of natural selection.

Von Neumann himself clearly acknowledged that this was the case for his cellular model. An extended discussion appears in (von Neumann 1966b, Sections 1.7, 1.8). There he explicitly accepted that any substantive interaction between two of his A-reproducers would be likely to cause "an unforeseeable class of malfunctions ... corrupting all reproduction" (p. 129), and that a similar result could be expected if the surrounding space for an A-reproducer were not initially quiescent (p. 130); and he did elaborate *ad hoc* methods whereby all such interactions could be avoided, such that descendents "will be distinct and non-interfering entities" (p. 127). He did, separately and briefly, suggest that Darwinian evolution could be "considered" in the context of his models, but then admitted that "the conditions under which it can be effective here may be quite complicated ones" (p. 131); with the benefit of hindsight this now appears to have been something of an understatement.

I do not claim that these various A-systems cannot support genuinely robust or viable A-reproducers of *any* sort (though I do *suspect* this to be the case). We should perhaps distinguish here two issues: the "robustness" of the A-parts, and the robustness of the A-machines composed of these parts. My own view, for what it is worth (and I conjecture that this was also von Neumann's view) is that the design of satisfactory A-parts is an almost trivial problem: the *difficult* thing is to organise these into complex, coherent, entities which can protect their own integrity in more or less hostile environments. Von Neumann solved (or, at least, showed the possibility of solving) the problem of how such complex A-machines could reproduce; and, in particular, how they could reproduce in a manner which would support the possibility of a Darwinian growth of A-complexity. He did *not* solve what is, in its way, a *prior* problem: that of how such A-machines could sustain themselves at all. This is

[7] I shall continue to refer to "natural" selection, even within "artificial" systems, consistent with the abstract interpretation discussed in (McMullin 1992a).

what I am calling the problem of *autonomy*; and I venture to suggest that it is much the harder problem.

6.3.1.2 Digression: The VENUS System

I shall digress here to discuss the VENUS system(s) described by Rasmussen *et al.* (1990). Technically, VENUS is the name for a simulator of one specific example of a more general class of A-system, which Rasmussen *et al.* refer to as *Coreworlds*. However, for convenience in what follows I shall use VENUS to refer loosely to both the simulator proper and the Coreworld which it simulates.

The VENUS Coreworld consists of an array of cells or memory locations (the "Core") in which reside instructions taken from a specified instruction set (Red Code), which is somewhat reminiscent of the instruction set of a simple modern computer. Instruction pointers, or virtual execution units, can execute these instructions. Instruction pointers may be dynamically created and destroyed (subject to a fixed maximum). Execution of any given instruction can freely affect other memory locations within some fixed radius. Execution uses up resources, which are replenished at a fixed rate; if insufficient resources are available for a given instruction pointer to continue execution (typically due to the existence of too many other instruction pointers in the same general region) then the pointer will be destroyed. Various effects in VENUS are stochastic rather than strictly deterministic.

In VENUS there is no simple notion of what constitutes an A-machine; but roughly speaking, one or more instruction pointers, together with some associated segment of core containing particular instructions, may be regarded as an A-machine.

Rasmussen *et al.* exhibit a single A-reproducer which can be embedded in VENUS. This is based on an original design by Chip Wendell called MICE (Dewdney 1987). This does *not* have the von Neumann self-reproducing architecture. Instead it uses something more akin to reproduction by self-inspection. This can be coerced into the von Neumann framework by regarding an A-machine as its own A-descriptor. This is feasible in the simple one-dimensional VENUS. It suffers by comparison to the more general von Neumann model in that it does not allow any flexibility in the genetic network.[8] Nonetheless, in the particular case of VENUS, it seems clear that the space of A-machines (which is to say A-descriptors) will, in fact, include a subspace of A-reproducers, derived from the MICE A-reproducer, which are "close" to each other under a reasonable interpretation of A-mutation. That is, it seems likely that VENUS does allow a solution to P_v, though only weakly following von Neumann's schema.

The advantage of VENUS over the other A-systems mentioned above is that, as a result of the relatively greater complexity of the individual cells, the simplicity of the geometry of the cellular space, and the relatively simplified (non-genetic) scheme of self-reproduction proposed, the basic self-reproducing A-machine is quite small—occupying only eight cells (memory locations, or A-parts). Empirical investigation of VENUS is thus quite feasible and it is precisely the results of one such investigation which are reported in (Rasmussen *et al.* 1990).

For my purposes the key result is this: the simple A-reproducer (MICE) described above was *not* viable. If VENUS is seeded with a single instance of this A-reproducer the population initially expands rapidly, but then these offspring interfere with and corrupt each other, leading the population to become extinct and/or sterile. In none of the tests reported did self-reproducing behaviour survive this initial transient. This directly illustrates and supports my claim that, surely, the same fate would befall the vastly more complex and fragile A-reproducers proposed by von Neumann, Burks, Thatcher, etc.

6.3.1.3 P_a Restated

The problem P_a may thus be restated as follows: we wish to exhibit an A-system which still retains the positive features which allowed a solution of P_v—the restriction to a "small" set of "simple" A-parts, the existence (in principle at least) of a set of A-reproducers spanning a wide range of A-complexity, connected under A-mutation, etc.—but which *additionally* satisfies a requirement that at least some of these A-reproducers (a subset still spanning a wide range of A-complexity) should be able to establish viable populations in the face of "reasonable" environmental perturbations, including, at the very least, fairly arbitrary interactions with other A-reproducers. That is, we should like to see natural selection occurring (rather than the A-reproducers being artificially prevented from interacting with each other, or simply going extinct). I shall refer to A-reproducers satisfying these conditions (if any) as *A-organisms*.

P_a does not have quite the crisp and explicit motivation which von Neumann was able to cite for P_v (the apparent *paradox* of evolutionary growth of bi-

[8] In particular, we cannot directly introduce the idea I have elsewhere called *Genetic Relativism* (McMullin 1992e, Section 4.2.6).

ological complexity). Nonetheless, I think it is clear that P_a is a good and interesting problem, and we could learn very much even from partial solutions of it. As I have mentioned, I also think it a very hard problem; but of course, we learn very little from the solution of easy problems.

As with P_v before it, P_a is not strictly formalisable; it relies particularly on an informal notion of what would represent "reasonable" environmental perturbation. And of course, I must emphasise yet again that, even if P_a could be solved more or less satisfactorily, it would not, in itself, mean that we could yet exhibit a Darwinian growth of A-complexity (or A-knowledge) in an artificial system: *that* would rely (among other things) on a correlation between S-value and A-complexity. But a solution to P_a would surely give us a vehicle for the investigation of this deeper and more fundamental issue: for Darwinian natural selection is precisely our best known example of a selective process having this characteristic—or, at least, so we conjecture.

P_a is well known in various forms; it might even be said to subsume all the problems of biological organisation, not to mention the problems of cybernetics, robotics, or even Engineering and Technology as a whole. More particularly, it is closely related to the problem of what Packard (1989) calls *intrinsic adaptation*. Similarly, Farmer & d'A. Belin (1992) have explicitly identified P_a (or at least something very much like it) as "probably the central problem in the study of Artificial Life".

I do not, of course, pretend to solve P_a; my intention is simply to leave it exposed as a kind of bedrock that underlies other things I have discussed. Indeed, in its way, P_a may be almost coextensive with the entire problem of Artificial Knowledge and its growth. For what distinguishes an A-organism from an A-reproducer—its autonomous ability to survive in a more or less hostile world, a world lacking any "pre-established harmony" (Popper & Eccles 1977, p. 184)—is precisely what I refer to as its A(rtificial)-knowledge; and what P_a demands is that we exhibit an A-reproducer with "enough" *initial* A-knowledge to allow at least the *possibility* for A-knowledge to then show further spontaneous, and open-ended, evolutionary growth.

I think that ongoing misunderstanding of von Neumann' original problem, and its solution, may have inhibited work on P_a somewhat; but there have, nonetheless, been various experiments and theories which may be said to have, deliberately or otherwise, addressed P_a. The following sections will be concerned with a critical review of a selection of these. I shall suggest that there has been some progress, but that it is still of a very limited kind.

6.3.2 The Genetic Algorithm

Burks explicitly identified John Holland as continuing von Neumann's work relating to evolutionary (Darwinian) processes in automata systems (Burks 1970b, p. xxiv). We may suppose therefore that Holland's work would be likely to address P_a. In fact, Holland has developed a number of quite distinct lines of enquiry in this general field; but that with which he is most closely identified is the idea of the so-called *Genetic Algorithm* (Holland 1975), and this section will be devoted exclusively to consideration of it.

"Genetic Algorithms" now come in many varieties, but I shall nonetheless refer simply to "the" Genetic Algorithm, to encompass all those variants which are more or less closely modelled upon, and largely derive their theoretical inspiration from, Holland's original formulation.

To anticipate my conclusion: it seems to me that the problem Holland sought to solve with the Genetic Algorithm is essentially disjoint from my P_a; it will follow (more or less) that, while the Genetic Algorithm may (or may not) be successful in solving its own problem, it can be discounted as offering any solution to P_a. Thus, I review the Genetic Algorithm, not to criticise it, but to clarify that it *is* irrelevant to my purposes. This is necessary as appearances might otherwise be deceptive: as noted, Burks specifically identified Holland as continuing von Neumann's programme; and Holland's work does, in some sense, involve the artificial realisation of processes of biological evolution. Without quibbling over words, I want to establish that the aspects of biological evolution preserved in the Genetic Algorithm are not those which are directly relevant to P_a.

I have reviewed the underlying philosophical commitments of Holland and his colleagues (Holland *et al.* 1986) elsewhere (McMullin 1992e, Section 3.8.3). I concluded there that processes which Holland *et al.* describe as *inductive* are, precisely, processes of *unjustified variation* in the sense of Campbell 1974a, 1974b); but I quite accept that, in given circumstances, some such processes may do "better" than others (in the sense of generating conjectures which are "biased" toward the truth). The formulation and comparison of processes in this respect is what Holland *et al.* call the "problem of induction", and what I shall refer to as *Holland's problem* or P_h—and I recognise it as a genuine and difficult problem.

Now, I contend that *von Neumann's* Problem, P_v, may be viewed as a special case of P_h: it is, precisely, P_h applied to the case of the growth of (inate) knowledge by Darwinian processes (whether in natural or artificial systems).

More specifically, P_v might be restated as follows. In order for A-complexity (A-knowledge) to grow by Darwinian means there must be a process (A-mutation) whereby A-reproducers of greater A-complexity can spontaneously arise from parents of lesser A-complexity. *Prima facie,* this is virtually inconceivable. It is difficult enough to see how a complex A-machine can successfully reproduce at all; but given that some can, we certainly expect these to be very much the exception rather than the rule. That is, if we think of A-machines as being identified with points in a space of "possible" A-machines, then we expect the A-reproducers to be extremely sparse in this space. Assuming that some such space will adequately represent the relationships between A-machines under any particular process of variation, then the very low (average) density of A-reproducers in the space seems to suggest that the possibility of a variation in any one A-reproducer giving rise even to another A-reproducer (never mind one of greater A-complexity) must be quite negligible.

Von Neumann's schema solves P_v essentially by pointing out that, via an A-reproducer architecture based on the use of a "genetic" (i.e. *programmable*) constructor, one can *decouple* the geometry of a variational space of A-reproducers from all the peculiarities of the particular A-parts etc. in use. Once this is done, it becomes almost a trivial matter to exhibit a space (which, in effect, characterises some process of spontaneous variation) with the property that, although the A-reproducers may still be rather sparse *on average*, they are concentrated into a very small subspace so that the density is locally high. Which is a roundabout way of saying that the spontaneous transformation of one A-reproducer into another A-reproducer (as opposed to a transformation into another A-machine which is *not* an A-reproducer) is quite possible—perhaps even "likely".

The key insight here is that the von Neumann self-reproducing architecture, based on reasonably "powerful" genetic machines, allows such a decoupling; it allows a "designer" space as it were, which can be so-configured that A-reproducers are "close" together. Indeed, once this self-reproducing architecture is proposed, it almost becomes difficult to see how the A-reproducers could *fail* to be close to each other in the relevant variational space (i.e. the space of dashed A-descriptors).

Granted, von Neumann himself never quite expressed matters in this way. However, he certainly recognised that the use of A-descriptors (i.e. the use of a fairly sharp genotype/phenotype decomposition) in his self-reproducing architecture was very important (von Neumann 1966a, p. 84; 1966b, pp. 122–123). In any case, regardless of his intentions, the fact remains that his schema solves a most substantive element of P_h (as interpreted in the context of Darwinian evolution).

We may say that P_h is still not "completely" solved of course. Von Neumann shows us firstly (and crucially) how a more or less arbitrary variational network or space can be overlaid on a set of A-machines; and he shows, secondly, a particular way of doing this such that set(s) of A-reproducers can be identified whose elements are "close" to each other. While this allows us to say that a given A-reproducer can plausibly be transformed into other, distinct, A-reproducers, it says nothing about the plausibility of such transformations resulting in increased A-complexity. If we think (*very* informally) of some measure of A-complexity being superimposed on the genetic space we may expect that, even still, the A-reproducers of "high" A-complexity may be very sparse in the space; so that it may seem that the likelihood of variations yielding increased A-complexity would still be quite negligible.

That this is precisely the point at issue in the Genetic Algorithm is emphasised by other elements of the problem situation which underlay Holland's work. The general notion of using vaguely "Darwinian" processes to achieve the growth of artificial knowledge had already received substantive prior investigation, but with mediocre results (e.g. Friedberg 1958; Friedberg *et al.* 1959; Fogel *et al.* 1966). While Friedberg *et al.* were commendably honest about this, Fogel *et al.* were, perhaps, less forthright. Lindsay's review of the work of Fogel *et al.* (Lindsay 1968) was harshly critical, and was arguably responsible for the virtual abandonment of any "Darwinian" approach for several years. Lindsay explicitly attributed the failure of such approaches to the relative sparsity of entities of high complexity in the relevant spaces.

Now one possible way of tackling this problem would be to try to handcraft the genetic space even further (beyond what had been explained by von Neumann), so that A-reproducers of "high" A-complexity *would* be dense, in at least some regions. This seems rather to beg the question however, for it effectively asks the designer to already know the relative complexities of all the A-reproducers involved. An alternative approach is to ask for more sophisticated procedures for negotiating this space (which is assumed to be given, and *not* to have A-reproducers of "high" A-complexity already conveniently packed closely together), than the simple, purely local, transformations implied by the notion of A-mutation as so far discussed. In my view, this is precisely what is being attempted with the idea of the Genetic Algorithm.

However: the crucial point is that none of this—neither von Neumann's solution of the original P_v, nor Holland's solution (if solution it be) of the enhanced form of P_v represented by P_h—addresses the core issue of *autonomy*. It is for this reason that I discount the Genetic Algorithm as offering any help in addressing P_a.

This argument does not quite make P_h and P_a *disjoint*. In particular, it does not necessarily mean that the Genetic Algorithm is, as I claim, *completely* irrelevant to P_a. The Genetic Algorithm *is* inspired by certain aspects of biological evolution; so, notwithstanding the fact that it was not formulated with P_a in mind, it (or at least its applications) might still address P_a to some extent. However, this is now a relatively minor issue and I shall not pursue it further here (for further discussion, see McMullin 1992e, Section 4.3.2).

6.3.3 Constraining the Interactions: The Tierra System

One strategy for addressing P_a is to consider A-systems which are more or less tightly constrained in the kinds of interactions allowed between A-machines. In this way it may be possible to guarantee that at least some of these will be viable, despite allowing interactions between them. Some work has been done along these lines (though perhaps not consciously with this end in mind) and I shall briefly review it here.

In the most extreme case, interactions between A-reproducers and their environment (or, more particularly, each other) can be effectively eliminated. This will certainly allow the A-reproducers to be "viable". As already discussed, von Neumann's original scheme for sustained self-reproducing activity was of this sort. Similar concepts were subsequently proposed by Laing (1975) and Langton (1986). But, as already mentioned, this simply sidesteps rather than solves P_a; indeed, once variation is allowed at all, it is virtually certain that the variant A-reproducers will no longer stay isolated from each other, and that all self-reproducing activity will quickly be destroyed.

The A-system proposed by Packard (1989) represents a more or less minimal retreat from this position. His set of A-reproducers ("bugs") are loosely modelled on the gross functionality of chemotactic bacteria. They have a fixed genetic structure consisting of just two genes, determining, respectively, their "food" threshold for undergoing reproduction, and the number of offspring resulting from a single act of reproduction. Other than these two characteristics all bugs are identical. Bugs exist in a two dimensional environment. *No* direct interactions between bugs are allowed—only indirect interactions via food consumption.

Due to the severely circumscribed interactions or perturbations between bugs and their environment they are generally more or less viable; but the allowed interaction is, indeed, sufficient to allow a minimal degree of (natural) selection. For the same reason, however, the possibility for A-knowledge to grow in this A-system is also severely impoverished. Natural selection can occur—but its effect is limited to, at best, selecting a combination of the food threshold for reproduction and number of offspring which is best matched to the characteristics of the available food supply. We may say that, through the evolution of the system, bugs (or, at least, bug-lineages) can, indeed, grow in their A-knowledge of their environment. But this is achieved at a cost of limiting the scope for such growth to a point where it is barely significant. In effect, Packard introduces natural selection only by abandoning von Neumann's achievement in the original solution of P_v—namely, the availability of a set of A-reproducers spanning an essentially infinite range of A-complexity (A-knowledge).

Packard of course recognises this limitation; indeed, it was a deliberate decision to attempt, initially, to design a *minimal* A-system which would exhibit natural selection. He explicitly notes the desirability of enhancing his A-system to include "a space of individuals that is open, in the sense that, as individuals change, they could have an infinite variety of possibilities" (Packard 1989, p. 154); if this corresponds to my requirement for an infinite range of A-complexity (or A-knowledge), then it identifies Packard's problem with P_a. In any case, the point is that, for the moment at least, Packard is still stating the problem rather than offering a solution.

Rizki & Conrad (1985) had earlier presented a much more sophisticated A-system (`Evolve III`), but in essentially the same genre. The range of A-complexity or A-knowledge is substantially wider, parameterised by fifteen distinct "phenotypic traits". The genotype/phenotype mapping is subject to a degree of variation also. Again, "genuine" natural selection can be achieved in this A-system, but the range of A-complexity or A-knowledge is still so sharply constrained that the scope for sustained growth of A-knowledge is unsatisfactory. The RAM A-system of Taylor *et al.* (1989) is a more recent, and independent development, but seems to share essentially the same strengths and weaknesses.

The final system which I wish to discuss here is the Tierra system described by Ray (1992).

Tierra can roughly viewed as a development of the VENUS system discussed in section 6.3.1 above—but with several fundamental modifications. Most importantly in the current context, Tierra involves the imposition of special constraints on the interactions between A-machines. In particular, a form of "memory protection" is introduced, which prevents the memory segment(s) "owned" by a given A-machine being perturbed by other A-machines. This now allows A-reproducers to be viable, but on its own actually makes them "too" viable—they become *invulnerable.* Thus, a single seed A-reproducer would quickly produce a population which exhausts the available memory, but there would be virtually no further activity; all the A-reproducers would be, in a certain rather strained sense, "alive"; but they could not function in any meaningful way.

To offset this, Ray introduces an automatic mechanism for killing A-machines (destroying instruction pointers and deallocating memory) so as to guarantee that a pool of unallocated memory is maintained which, in turn, ensures the possibility of continuing activity. Very roughly speaking, this is a "mortality" mechanism, operating on a FIFO basis—the "older" an A-machine is, the more likely that it will be killed in this way—though there are other factors which may qualify this to a limited extent.

Tierra differs from VENUS in a variety of other respects also. For example, the process scheduling rules in Tierra are rather simpler than in VENUS. More substantively, although Ray continues to use a form of self-reproduction based on self-inspection (rather than a properly genetic system in the von Neumann sense), his instruction set (Tierran) is quite different from the Red Code of VENUS. Ray argues that Tierran should exhibit enhanced "evolvability" compared to Red Code. In my terms, Ray is compensating for the inflexibility associated with reproduction by self-inspection by attempting to directly handcraft the "phenotype" space. This is a perfectly reasonable strategy; but again, it would seem preferable to allow for full blown *Genetic Relativism* (McMullin 1992e, Section 4.2.6) instead. In any case, although Ray places significant emphasis on the differences between Tierran and Red Code, it is difficult to assess his claims in this regard: he does *not* present any empirical test of the specific hypothesis that Tierran has improved "evolvability" compared to Red Code (which would involve presenting a comparison of systems in which the instruction set is the *only* difference between them). My own conjecture (equally untested) is that the instruction set is of relatively little significance; the *crucial* difference between VENUS and Tierra is, in my view, the use of memory protection and controlled mortality.

Unlike VENUS, self-reproduction behaviour in Tierra can generally persist for indefinitely long periods of time. This is a direct consequence of the memory protection and controlled mortality mechanisms. As a result, Ray's empirical investigation of Tierra *has* demonstrated what I regard as sustained Darwinian evolutionary processes, including some rather dramatic phenomena. In particular, Ray has exhibited the emergence of various kinds of *parasitism.* That is, A-reproducers emerge which partially exploit code, and possibly even instruction pointers, owned by other A-reproducers, in order to complete their own reproduction. Ray (1991) has also reported the emergence of A-reproducers in which more or less "complex" optimizations of the reproduction mechanism have occurred.

Thus, A-knowledge has indeed grown in Tierra, by Darwinian mechanisms. We may reasonably say, for example, that a basic parasite "knows" (or at least "expects") that certain other A-reproducers will be present in its environment, with which it can interact in certain ways in order to complete its reproduction. Similarly, A-reproducers exhibiting immunity to certain kinds of parasitism may be said to "know" about those kinds of parasitism. The optimization of the reproductive mechanism, mentioned above, involves "knowing" about certain aspects of the underlying process scheduling mechanism (namely that "bigger" A-machines get allocated more CPU time than "smaller" ones).

These are all substantive results. Tierra is a definite improvement on the other A-systems considered in this section, in that the space of A-reproducers is once again very large and diverse, as it was in the original von Neumann proposal. Tierra is also an improvement over the von Neumann proposal (and its close relatives) in that at least some A-reproducers are viable, despite interactions between them, and natural selection can indeed be exhibited as a result. In my view, Tierra represents the best example to date of something approximating Artificial Darwinism.

On the other hand, Tierra can hardly be said to seriously confront P_a. A `Tierran` A-machine is not, by and large, responsible for its own integrity—that is essentially guaranteed by the memory protection mechanism; so the difficulties represented by P_a are not directly addressed within `Tierra` (as it stands). In this sense, the potential for the growth of A-knowledge in `Tierra` would seem to be strictly limited. This suspicion is borne out, at least by the results so far; while there has certainly been *some* interesting, and even surprising, growth of A-knowledge in my terms, it still seems to have been very limited, being concerned almost exclusively with fine tuning of reproductive efficiency. I suggest that this will continue to be the case, as long as the substance of P_a is effectively bypassed. Indeed, I may annunciate the following crude, but general, principle: the stronger are the constraints on interactions by A-reproducers (which is to say the weaker the attack on P_a) then the smaller must be the scope for A-knowledge to be the subject of natural selection—for it is only by mediating interaction that A-knowledge can attain a selective value. In `Tierra`, of course, the constraints on interaction are very strong indeed.

6.3.4 Autopoiesis: The Organisation of the Living?

> . . . the process by which a unity maintains itself is fundamentally different from the process by which it can duplicate itself in some form or another. Production does not entail reproduction, but reproduction does entail some form of self-maintenance or identity. In the case of von Neumann, Conway, and Eigen, the question of the identity or self-maintenance of the unities they observe in the process of reproducing and evolving is left aside and taken for granted; it is not the question these authors are asking at all.
>
> Varela (1979, p. 22)

The path I have presented thus far, to the recognition of the problem of autonomy, P_a, is a somewhat tortuous one, proceeding via the failure of von Neumann style "self-reproducing automata" to actually support a Darwinian, evolutionary, growth of complexity (or knowledge). There is an alternative, arguably more direct, route which has been pioneered by Humberto Maturana and Francisco Varela (Maturana & Varela 1980; Varela 1979).

Briefly, the difficulty with the von Neumann A-reproducers can be stated in this way: they are, evidently, "unities" only by convention, relative to us as observers—they do not assert or enforce their own unity within their domain of interactions. In fact, this is true of what we typically call "machines" or "automata" in general, and is a crucial difference between such systems and those systems which we call "living". This is, perhaps, clear enough on an intuitive level, but it is quite another matter to elaborate exactly what this distinction consists in—what does it mean for an entity to "assert" its unity. This is the problem which Maturana & Varela have tackled; and we can now see that it is a problem in its own right, which is actually logically *prior* to von Neumann's problem of the growth of automaton complexity (by Darwinian evolution), as it queries what we should regard as an "automaton" in the first place. The solution which Maturana & Varela propose is this: what distinguishes "living" or properly "autonomous" systems is that they are *autopoietic*. This is defined as follows:

> The authors [Maturana & Varela 1973b] first of all say that an autopoietic system is a homeostat. We already know what that is: a device for holding a critical systemic variable within physiological limits. They go on to the definitive point: in the case of autopoietic homeostasis, the critical variable is *the system's own organization*. It does not matter, it seems, whether every measurable property of that organizational structure changes utterly in the system's process of continuing adaptation. *It* survives.
>
> Beer (1973, p. 66, original emphasis)

> The autopoietic organization is defined as a unity by a network of productions of components which (i) participate recursively in the same network of productions of components which produced these components, and (ii) realize the network of productions as a unity in the space in which the components exist. Consider for example the case of a cell: it is a network of chemical reactions which produce molecules such that (i) through their interactions generate and participate recursively in the same network of reactions which produced them, and (ii) realize the cell as a material unity. Thus the cell as a physical unity, topographically

> and operationally separable from the background, remains as such only insofar as this organization is continuously realized under permanent turnover of matter, regardless of its changes in form and specificity of its constituent chemical reactions.
>
> Varela *et al.* (1974)

Accepting, at least tentatively, this vision of what would properly constitute an "autonomous" system, my "problem of autonomy" (P_a) can now be recast in a somewhat more definite form: can we exhibit an A-system which still retains the positive features which allowed a solution of P_v—the restriction to a "small" set of "simple" A-parts, the existence (in principle at least) of a set of A-reproducers spanning a wide range of A-complexity, connected under A-mutation, etc.—but which *additionally* satisfies a requirement that these A-reproducers should be *autopoietic* unities?

As far as I am aware, this problem has not been previously explicitly formulated, much less solved. However, a simpler problem *has* been previously tackled and solved: this is the problem of exhibiting an A-system which can support autopoietic (autonomous) A-machines of *any* kind. The original solution was presented by Varela, Maturana & Uribe (1974), and further developments have been reported by Zeleny (1977) and Zelany & Pierre (1976). This work is also reviewed in (Varela 1979, Chapter 3).

The A-systems described by these workers were inspired to an extent by the work of von Neumann, and bear some similarity to two dimensional cellular automata. However, these A-systems are also very distinctive as a result of being deliberately designed to support autopoietic organisation. In any case, I shall not present a detailed description here. The essential point, for my purposes, is that the possibility of exhibiting artificial autopoietic unities within a suitable A-system has been satisfactorily demonstrated; indeed, Zeleny (1977) has indicated that a primitive form of *self-reproduction* of such autopoietic entities may be demonstrated (though I should emphasise that this bears no significant similarity to the *genetic* self-reproduction envisaged by von Neumann; this illustrates yet again the shallowness of the idea that von Neumann worked on "the" problem of self-reproduction as such).

It thus seems that the two aspects of my P_a have been *separately* addressed, successfully, within the general framework of (two dimensional) cellular automata. That is, von Neumann and his successors have shown how A-reproducers can be organized such that there will exist an A-mutational network linking low complexity A-reproducers with high complexity A-reproducers, using the idea of "genetic" A-descriptors; and Varela, Maturana, and others, have shown how properly robust or *autonomous* A-machines (and even A-reproducers of a kind) can be organized. P_a calls for both these things to be exhibited at once. The separate results certainly suggest that the general cellular automata framework is rich enough or powerful enough to allow a solution of P_a.

As far as I am aware, however, no one has yet explicitly attempted this synthesis—and the difficulty of achieving it should not be underestimated. In the first place, the A-systems which have yielded these separate results bear only very limited similarities. More importantly, the A-machines under consideration, embedded in these distinct A-systems, are radically different *kinds* of entity. Whereas an instance of one of von Neumann's original A-machines can be reasonably well defined simply by identifying a fixed core set of cells (A-parts) which constitute it, the autopoietic A-machines of Varela *et al.* can potentially retain their unity or identity even through the replacement of all of their A-parts.

This last point actually suggests the possibility of a radical reinterpretation of some of the A-systems already discussed previously, particularly `VENUS` and `Tierra`. While it is clear that the entities which are *conventionally* regarded as the A-machines in these systems (namely, the code fragments associated with a single virtual CPU) are *not* autopoietic, it seems possible that certain aggregations of these *may* be validly said to realise a primitive (or *partial*) autopoietic organisation. For example, it seems that this may be an alternative, and potentially fruitful, view of the emergence of what Rasmussen *et al.* (1990, p. 119) actually call "organisms" in the `VENUS` system; and, equally, this may be a valid view of the phenomena which Ray (1992) describes in terms of the emergence of "sociality" in the `Tierra` system. But of course, if this alternative view is adopted, then the "higher-level", autopoietic, A-machines now being studied are no longer typically self-reproducing in any sense, never mind being self-reproducing in the von Neumann, genetic, sense.

Thus, it is clear that, while the work on artificial autopoiesis yields a considerable and valuable clarification of P_a, and perhaps even some progress toward its solution, it is not yet a solution as such.

6.3.5 The Holland α-Universes

I now finally turn to what is, superficially at least, a quite different strategy for tackling P_a. Insofar as the problem has been explicitly tackled up to this point, the typical approach has been to attempt to handcraft at least one initial robust or viable A-reproducer. In practice this has been effective only if the environmental perturbations are made almost negligible (such as in the case of the `Tierra` system). In this way a superficial "viability" can be achieved, but without actually realising *autonomy*, in the autopoietic sense, at all; which is to say, P_a is being avoided rather than solved. In itself this is unsurprising. We already know that even relatively simple biological organisms are much more complex that the most complex extant technology. The question is how to bridge this gap (assuming that to be even possible!).

One obvious suggestion is that we should take a further lesson from the biological world (i.e. in addition to, or perhaps going beyond, the central idea of Darwinian evolution). We know, or at least presume, that biological organisms arose by some kind of spontaneous process from a prior, *abiotic,* environment; so a possible strategy for the development of artificial "organisms" (in the sense of entities which satisfy the conditions for a solution of P_a) may be to see if *they* might spontaneously arise in an artificial, abiotic, environment. That is to say, instead of attempting to directly construct artificial life, we attempt to realise an artificial version of the original *genesis* of life.

As it happens, a proposal of essentially this sort was made some years ago (albeit for somewhat different reasons) by John Holland, in the form of what he called the α-Universes (Holland 1976). The system proposed by Holland (which I denote α_0) bears some resemblance to the `VENUS` and `Tierra` systems, involving a one dimensional space, supporting putatively self-reproducing A-machines. However, it differs in several important respects: it has an overtly bio-chemical inspiration; the putative A-reproducers use a von Neumann style *genetic* mechanism; and, or course, the A-reproducers are expected to be capable of spontaneous emergence.

Holland provided some initial theoretical analysis of his proposal, but he then left the idea aside. α_0 has recently been reexamined, including a comprehensive programme of empirical testing (McMullin 1992d; 1992e, Chapter 5). I shall not detail the results of that investigation again: it is sufficient to note the conclusion, which is that α_0 does *not* yet provide any substantive advance toward a solution of P_a. The A-reproducers in α_0, such as they are, are just as fragile as in, say, `VENUS`. α_0 does not provide any prospect for the spontaneous emergence of *robust* A-reproducers, and does not, therefore, provide a basis for the solution of P_a.

6.4 Conclusion

I do not, of course, *know* how one might best proceed in the light of the what has been presented here; but there are two distinct avenues which seem to me worth considering further.

Firstly, it seems that at least one part of the deficiency of α_0 hinges on the fact that von Neumann style reproduction involves *copying* and *decoding* an information carrier, where the decoding must be such as to generate (at least) a copy of the required copying and decoding "machinery". α_0 fails to sustain this kind of behaviour because (*inter alia*) the maximum information capacity of its carriers (in the face of the various sources of disruption) seems to be of the order of perhaps 10 bits, which is insufficient to code for any worthwhile machinery—even the relatively simple copying and decoding machinery constructible in α_0.

A more plausible model for the spontaneous emergence of properly genetic A-reproducers *might* therefore involve a universe in which certain information carriers, of capacity (say) an order of magnitude larger than that required to code for minimal decoding machinery (in the particular universe), can be copied *without any specialised machinery at all.* In such a system there may be potential for a Darwinian evolutionary process to begin more or less immediately, in which more sophisticated phenotypic properties might, incrementally, become associated with the information carriers—possibly then culminating in a full blown "decoding" (or embryology).

This is, of course, rather speculative; but, as it happens, it is closely related to a general model for the origin of *terrestrial* life which has been championed by Cairns-Smith (1982). This is based on *inorganic* information carriers, which could conceivably be replicated without the relatively complex apparatus required for RNA or DNA replication. It seems to me that it would now be a promising research program to adopt Holland's original *strategy* (which is to design relatively simplified model chemistries, loosely based on cellular automata, in which to examine the origin of "life"), but to replace his detailed models (the α-Universes) with models based on different theoretical considerations—such as those of Cairns-Smith.

The second avenue I can envisage for challenging the limitations of α_0 turns on a point which is both subtle and fundamental. I have already anticipated this issue in section 6.3.4 above.

Briefly, the situation is this. As long as we consider an instance of an A-machine (in, say, α_0, or VENUS, or Tierra) as corresponding to a particular, fixed, set of A-parts, then it makes sense to regard the mutually recursive relations of production between these A-parts as realising a form of self-reproduction—such a set of A-parts is (in principle at least) capable of bringing new and separate instances of such sets into existence.

But this is not the only possible way of looking at things. We could, instead, regard an A-machine as corresponding to the set of recursive relations of production rather than a particular set of A-parts which happen to realise these relations. These relations of production are then recognised as being partially autopoietic: such an entity is (or, at least, could be) capable of sustaining itself, by virtue of this organisation, despite turnover of some or all of its constituent A-parts.

From this perspective, fundamentally related phenomena can now be recognised as occurring in these distinct systems (α_0, VENUS, and Tierra). I have talked very loosely in terms of A-reproducers as being potentially "robust" or "viable"; but the fact is that, as long as by "A-reproducer" I meant a single fixed set of A-parts, there was never any possibility of their being "autonomous" in the strong sense of being *autopoietic.* As it happens, the putative α_0 A-reproducers turned out not to be "viable" anyway (just like the A-machine MICE in VENUS); but, even if they had been "viable", it seems that it could only have been, at best, the cosseted "viability" of the A-reproducers in Tierra with their inviolable memory allocations. By definition, no *static* set of A-parts (structures) in α_0 can realise the *dynamic* homeostasis of its own identity, which would be characteristic of properly autopoietic viability or autonomy.

By contrast, if we turn our attention to "populations" of structures in α_0—the equivalent of considering "organisms" in VENUS or "sociality" in Tierra—we *can* encounter the possibility of at least partially autopoietic organisation. Granted, in α_0 as it stands, the autopoiesis is not effective—such populations actually die out—but (with the example of Tierra before us) we may anticipate that some modified α-Universe could overcome this. The point is that the kinds of entities which we might properly regard as autonomous are *not* the kinds of entities which could be regarded as self-reproducing; and, moreover, the "higher level", properly autonomous entities, are not, in general, self-reproducing in any sense, and are *certainly* not genetically self-reproducing in the von Neumann sense of permitting an open-ended growth in complexity.

Can we envisage a path toward making the properly autonomous entities ("organisms" in VENUS, "social systems" of Tierra, "populations" in α_0) self-reproducing, in the von Neumann sense?

Well, the first point is that to have *any* kind of self-reproduction, we would probably need some mechanism for the formation and maintenance of *boundaries* by the autopoietic entities. Some kind of boundary formation is actually part of the definition of fully fledged autopoiesis. Furthermore, a boundary seems to be logically necessary if we wish to talk about self-reproduction: unless the entities establish well defined boundaries then it is entirely unclear what could possibly qualify as self-reproduction. In VENUS, Tierra, or α_0, as they stand, there are no such mechanisms for boundary formation (capable of bounding the *relevant* entities). Boundary formation has, of course, been exhibited in the A-systems pioneered by Varela *et al.* (1974). These systems, by contrast to VENUS, Tierra and α_0, are *two* dimensional rather than linear. On the other hand, the introduction of a kind of boundary mechanism has been previously outlined by Martinez (1979), in a modification of α_0 which would still be one-dimensional. Thus, while two-dimensionality is probably not essential here, it certainly provides conceptual simplification, and makes visualisation much easier.

Incidentally, it seems plausible that the introduction of an appropriate boundary mechanism could positively help in overcoming the primary deficiency of α_0 identified by the empirical testing, that even the putatively autopoietic populations cannot actually sustain themselves.

In any case, assuming the introduction of mechanisms allowing for the construction and maintenance of such boundaries, it is clear that self-reproducing autopoietic entities can be established, in the manner already described by Zeleny (1977). Briefly, once one has a bounded autopoietic entity of any sort then, since it already incorporates processes capable of reestablishing all its component relationships, it should be a relatively trivial matter to arrange for it to progressively grow *larger.* Once this is possible, then one need only add a mechanism for the boundary to rupture in such a way that it can be reformed into two closed parts, and a primitive form of self-reproduction is achieved. There seems no reason, in principle, why this general kind of process cannot be achieved in A-systems derived from the VENUS, Tierra or α_0 models.

Doing this based on the VENUS or Tierra models would yield a form of self-reproduction which might still be said to be *impoverished* in the sense that, insofar as "information carriers" are being reproduced, this is occurring by self-inspection, without any overt genotype/phenotype distinction, or von Neumann style *decoding*. Still, although I have arrived at this from a completely distinct direction, this idea actually corresponds rather closely to the first suggestion which I outlined in this section, following Cairns-Smith (1982), of arranging for the possible existence of reasonably high capacity "information carriers" which could be "reproduced" without the aid of any special or elaborate machinery. It may thus be a useful, and perhaps even essential, step toward more sophisticated self-reproduction techniques.

Conversely, if we used α_0 as our starting point, and succeeded in modifying it to support reproduction of bounded, autopoietic, "populations", then we would have entities which *do* exhibit a "von Neumann style decoding"; but, of course, they would be impoverished in a different manner, namely that the functionality available in α_0 is extremely impoverished anyway and there certainly could not exist a space of such autopoietic A-reproducers which would span a wide range of A-complexity.

This is all rather vague and informal, and I do not pretend that it has more than heuristic value. Nonetheless, it seems that there may be some limited grounds for optimism here. If the various phenomena which have been separately exhibited in this diverse range of A-systems can be consolidated into a single system, then it seems that some significant progress may then be possible in the solution of P_a.

Acknowledgements

This essay is part of an ongoing attack on the problem of realising the spontaneous growth of Artificial Knowledge by Darwinian (or any other!) means. I have benefited greatly from discussions with colleagues, particularly Noel Murphy in DCU, and John Kelly of University College Dublin. I am indebted to the School of Electronic Engineering in DCU (particularly through the agency of its Head, Charles McCorkell) for continuing encouragement, not to mention material support. All errors remain, of course, my own responsibility.