Chapter 4

Artificial Darwinism

4.1 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*:

> Artificial Life is the study of man-made systems that exhibit behaviors characteristic of natural living systems. It complements the traditional biological sciences concerned with the *analysis* of living organisms by attempting to *synthesize* life-like behaviors within computers and other artificial media. By extending the empirical foundation upon which biology is based *beyond* the carbon-chain life that has evolved on Earth, Artificial Life can contribute to theoretical biology by locating *life-as-we-know-it* within the larger picture of *life-as-it-could-be*.
>
> Langton (1989b, p. 1, original emphasis)

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 chapter will be concerned with a very selective review of work carried out by a small number of researchers. The choice of which work to highlight in this way is a personal one, but is not arbitrary. I shall concentrate almost exclusively on von Neumann's seminal investigations, which may be taken almost as having defined the field. 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 of the work 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was finally edited, completed, and posthumously published, together with a series of related lectures (also previously unpublished), under the general title *Theory of Self-Reproducing Automata* (Burks 1966d).

I say this chapter provides a "review" but it should perhaps be put a little more strongly than that. Briefly, my contention is that von Neumann's original work has been, at best, incompletely understood; and that (perhaps as a direct result) the research programme which he proposed has foundered. Thus, the 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. One such alternative strategy would be to abandon altogether the attempt to *create* Artificial Life (and thus Artificial Darwinism); for such creation *may* be simply too difficult, at least for the time being. Instead we might try to create an (artificial) system which is admitted to be devoid of "life", at least initially, but in which "life" may spontaneously arise. That is, we could redirect our attention away from the broad sweep of evolutionary biology (which "pre-supposes" the existence of life, albeit of a "primitive" kind), and concentrate instead on capturing the *genesis* of "life" in an artificial system. This may (or may not!) be a more tractable problem; in any case, it will subsequently become the specific concern of Chapter 5.

4.2 Von Neumann's *Theory of Automata*

4.2.1 Background

> You had to be a quick note-taker indeed if you were going to follow one of von Neumann's lectures. During his seminars (Fuld Hall's seminar room was right across the hallway from his office) he'd write dozens of equations on the blackboard, jamming them all into a two-foot square space off to one side. As soon as he was finished with one formula he'd zip it away with the eraser and replace it with another one. He'd do this again and again, one right after the other—an equation and *zzzip*, another one and *zzzip*—and before you knew it he'd be putting the eraser back on the ledge and brushing the chalk dust from his hands. "Proof by erasure," his listeners called it.
>
> Regis (1987, p. 104)

In the late 1940's John von Neumann began to develop what he intended as a truly *general* "theory of automata". By "automaton" von Neumann meant, roughly, any system which could be described or understood as a more or less "complex" whole made up of "simple" parts having prescribed properties. In other words, an automaton is any system which is amenable to a strictly reductionist analysis (or synthesis, for that matter). This is not to say that von Neumann was a "reductionist" in any general or cosmological sense (I do not know what, if any, metaphysical positions he adopted). The point is rather that the scope of his "theory of automata" was restricted, by definition, to just those systems which *are* reducible in this kind of operational sense.

The class of automata was, of course, to include artificial systems in general: after all, reductionist explanation is (for the time being at least) the *sine qua non* of all successful engineering; to this extent "the theory of automata" would almost be better called "the theory of engineering". But von Neumann also included biological systems (organisms in particular), at least tentatively; that is, to whatever extent biological phenomena may yield to a reductionist explanation (and this is ultimately, of course, an open question), then the study of these phenomena would fall properly within his theory of automata.

Von Neumann's automata theory thus involves two quite distinct kinds of question:

1. The characterisation of the "primitive" parts. In the simplest case von Neumann required that these should be what would now be called *finite state*

machines, of some sort. That is, any given primitive part would have some specified set of "inputs", some specified set of "outputs", and its "instantaneous" outputs would be determined by its instantaneous inputs and its instantaneous, internal, "state"—where inputs, outputs and internal states each admit of only finitely many distinguishable values.

2. The organisation of the parts into complex wholes, having some coherent properties and behaviours. In particular, certain sets of such complex wholes, defined in some way, may be identified as "automata" of some more or less interesting kind.

The two questions are clearly interrelated. Thus: the potential or scope for complex organisation must, in the final analysis, be constrained by the properties of the primitive parts; but conversely, we may speculate that certain "interesting" behaviours of a whole may be largely independent of the detailed properties of the parts, being chiefly a reflection of their *organisation*.

Von Neumann proposed, initially at least, to address questions of the first kind (the characterisation of primitive parts) by a process of unilateral *axiomatization*:

> Axiomatizing the behaviour of the elements means this: We assume that the elements have certain well-defined, outside, functional characteristics; that is, they are to be treated as "black boxes." They are viewed as automatisms, the inner structure of which need not be disclosed, but which are assumed to react to certain unambiguously defined stimuli, by certain unambiguously defined responses.
>
> This being understood, we may then investigate the larger organisms that can be built up from these elements, their structure, their functioning, the connections between the elements, and the general theoretical regularities that may be detectable in the complex syntheses of the organisms in question.
>
> I need not emphasize the limitations of this procedure. Investigations of this type may furnish evidence that the system of axioms used is convenient and, at least in its effects, similar to reality. They are, however, not the ideal method, and possibly not even a very effective method, to determine the validity of the axioms. Such determinations of validity belong primarily to the first part of the problem. Indeed they are essentially covered by the properly physiological (or chemical or physical-chemical) determinations of the nature and properties of the elements.
>
> von Neumann (1951, pp. 289–290)

The paper from which the above is quoted was originally read at the Hixon Symposium (on *Cerebral Mechanisms in Behavior*) in September 1948. Von Neumann returned again, and in more detail, to this issue during the following year,

in the course of a series of lectures delivered at the University of Illinois (finally published as von Neumann 1966a). He there concluded:

> ... while the choice [of the "elementary parts"] is enormously important and absolutely basic for the application of the axiomatic method, this choice is neither rigorously justifiable nor humanly unambiguously justifiable. All one can do is to try to submit a system which will stand up under common sense criteria.
>
> von Neumann (1966a, p. 77)

I emphasise von Neumann's stipulation of the essentially informal nature of the axiomatization procedure, because it underlines the *contingent* nature of his results—they are valid, and are claimed to be valid, only within the scope of certain specified axiomatisations. We shall see, in due course, that this important point has been overlooked by at least one subsequent worker (Frank Tipler—see section 4.2.4), who has then gone on to impute quite unjustified claims to von Neumann.

With regard to the second kind of question—the organisation of parts into complex wholes—von Neumann concentrated on one particular problem, which he identified roughly as the *growth of complexity*. More specifically, he wanted to establish that there is nothing fundamentally paradoxical about the notion of a complex automaton being able to construct another which is as complex as itself ("self-reproduction"—a prerequisite for natural selection—being the prototypical example); or, more substantially, about the notion of an automaton spontaneously becoming, via construction or otherwise, *more* complex. Together, these properties would permit, though not, of course, guarantee, the spontaneous growth of complexity via Darwinian evolution. He sought to do this by actually exhibiting these possibilities—automaton self-reproduction in a form supporting the *possibility* of spontaneous, heritable, growth in automaton complexity—within some particular, more or less "reasonable", axiomatization of "primitive parts" and "automaton".

In effect, von Neumann was interested in showing that certain conditions, which seem to be *necessary*, though not sufficient, for the spontaneous growth of complexity by Darwinian evolution, *can* be satisfied within relatively simple (reductionist) systems. This result would, of course, open up the prospect of

actually building artificial systems, computational or otherwise, which satisfy these minimal conditions.

In what follows I shall interpret von Neumann's informal notion of automaton *complexity* as being synonymous with what I have, in the previous chapter, called subjective *knowledge*, and I shall use the terms interchangably. That this interpretation is a reasonable one may, perhaps, be most clearly seen from the following passage:

> There is a concept which will be quite useful here, of which we have a certain intuitive idea, but which is vague, unscientific, and imperfect ... I know no adequate name for it, but it is best described by calling it "complication." It is effectivity in complication, or the potentiality to do things. I am not thinking about how involved the object is, but how involved its *purposive operations* are. In this sense, an object is of the highest degree of complexity if it can do very difficult and involved things.
>
> von Neumann (1966a, p. 78, emphasis added)

I review this notion of "complication" (and its rôle in biological Darwinism) in more detail in (McMullin 1992b, pp. 5–7). For the present purposes, it is sufficient to note that the problem of the growth of automaton "complexity" (in von Neumann's sense) is thus essentially equivalent to the problem of the growth of "knowledge" as I have discussed it heretofore. Von Neumann's work is therefore of very direct relevance to the concerns of this Thesis and deserves careful and detailed consideration.

Note carefully here that von Neumann's concern here was *not* with "self-reproduction" *per se*, but with the general problem of the construction of complex automata by other automata in such a way that complexity need not degenerate, and may even increase; and the reason for this concern was because of its relation to the fundamental problems of biological evolution. This must be emphasised because "self-reproduction" is a vague concept which admits of trivial as well as interesting interpretations, a fact of which von Neumann was keenly aware. He sought to avoid triviality in two ways. Firstly, he constrained what should be regarded as a "reasonable" axiomatization (specifically, constraining the powers of the primitive parts). But secondly, and crucially in my view, he constrained the phenomena which should be admitted as proper examples (for his purposes)

of self-reproduction. Both these points were covered in the 1949 lecture series, already mentioned above, delivered at the University of Illinois:

> ... one may define parts in such numbers, and each of them so large and involved, that one has defined the whole problem away. If you choose to define as elementary objects things which are analogous to whole living organisms, then you obviously have killed the problem, because you would have to attribute to these parts just those functions of the living organism which you would like to describe or to understand. So, by choosing the parts too large, by attributing too many and too complex functions to them, you lose the problem at the moment of defining it.
>
> von Neumann (1966a, p. 76)

> ... One of the difficulties in defining what one means by self-reproduction is that certain organizations, such as growing crystals, are self-reproductive by any naive definition of self-reproduction, yet nobody is willing to award them the distinction of being self-reproductive. *A way around this difficulty is to say that self-reproduction includes the ability to undergo inheritable mutations* as well as the ability to make another organism like the original.
>
> von Neumann (1966a, p. 86, emphasis added)

So it is clear that, insofar as von Neumann was interested in some "problem" of self-reproduction, it was, via the notion of inheritable mutations, purely in its rôle in the (Darwinian) growth of complexity.

Now, of course, these conditions, stipulated by von Neumann to avoid triviality, are not *formal*. Indeed, according to von Neumann, they are not even *formalisable*. I have already quoted him explicitly to the latter effect in the case of choosing an axiomatization. He implicitly makes the same point, though perhaps less strongly, in regard to the possibility of formalization of "inheritable mutation"; for this clearly refers to the possibility of "mutations" which may involve increased *complexity*, and, again as already quoted, von Neumann admits the vague and informal nature of his concept of complexity.

My reason for drawing out this point is that it seems to have been missed or obscured by at least some subsequent workers; in particular, there has been a perception that von Neumann was concerned with self-reproduction *as a problem in itself*. This is a part of what I shall call the *von Neumann myth*. This myth has had various negative effects, such as, for example, spawning an extended attempt to *formalise* a criterion for "non-trivial self-reproduction"—an attempt which I believe to have been unnecessary, confusing, and ultimately sterile (as

von Neumann clearly anticipated in the first place). To reiterate: my view is that von Neumann was not at all concerned with self-reproduction as a problem in itself (indeed, discussion on that basis can hardly *avoid* triviality); but rather with self-reproduction as a facet of a much more substantive problem—the growth of automata complexity (particularly via Darwinian evolution). I shall return to this issue in more detail in section 4.2.7 below.

Von Neumann's earliest expositions, in 1948/49 (first privately at the Princeton Institute for Advanced Studies, and then at the Hixon symposium, and later again in the lectures, already quoted, delivered at the University of Illinois) were in terms of a model which was very informal, but which sufficed to allow him to at least outline his arguments. Subsequently, he set out to provide a mathematically rigorous axiomatization, and derivation, of his results. He brought this to a fairly advanced stage in a manuscript written during 1952/53. The essential aspects of this model were presented in the Vanuxem Lectures delivered at Princeton University in March 1953. Von Neumann himself did not wish to write up these lectures separately from his manuscript; in the interim, it was arranged that John Kemeny write an article, based on the Vanuxem Lectures. This was published as (Kemeny 1955). In late 1953 von Neumann put his manuscript aside, unfinished; in the event, he was never to return to it. John von Neumann died in February 1957, after an extended illness.

Von Neumann's manuscript, tentatively entitled *The Theory of Automata: Construction, Reproduction, Homogeneity*, was finally edited and completed by A.W. Burks, and published posthumously as (von Neumann 1966b).

In the following sections I shall review von Neumann's work on the theory of automata in some detail. I adopt the following procedure. First, I restate, as clearly as possible, the particular problem which (I claim) von Neumann was setting out to solve (which I shall term P_v). Next I digress temporarily to discuss Turing's work on computing automata, in order to introduce certain ways in which von Neumann planned to exploit or generalise this work. Then I consider von Neumann's proposed solution to P_v. This involves initially *assuming* than some system or axiomatization of automata supports certain more or less plausible phenomena; this discussion corresponds essentially to von Neumann's

early, informal, presentations of his ideas, and represents what I call his *core* argument. This is followed by von Neumann's correction of a minor flaw in this core argument. I then review the demonstration(s) by von Neumann and others that the required phenomena *are*, in fact, supported in at least one particular axiomatization of automata (this is based on a cellular automaton formulation, and corresponds to von Neumann's later, unfinished, manuscript), and discuss the extent to which this successfully solves P_v. Having completed the presentation of von Neumann's solution, I present some mild criticism or clarification of it, showing how it can perhaps be strengthened in certain ways. I close this detailed discussion of von Neumann's work by returning, once again, to the question of what *problem* he was actually trying to solve. I distinguish sharply between my own view on this and the somewhat contrary views seemingly expressed by von Neumann himself, and by a number of other commentators.

4.2.2 Von Neumann's Problem (P_v)

> Among the many questions which our discussion of self-reproducing automata raises are 'Whence come the components out of which our automata are made?' and 'Given that such automata exist, how might one imagine them to evolve?' It is not our purpose in this section to answer these questions—would that we could—but rather to suggest some interesting avenues towards their solution.
>
> Arbib (1969a, p. 214)

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 *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 carefully, at this point, 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 may well 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 unjustified (in the sense of "unjustified variation" introduced in Chapter 3) 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 (initial) problem in the theory of automata, which I shall denote P_v, is 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

[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).

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 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*). This notion of A-complexity or A-knowledge is necessarily *in*formal; it will be interpreted in essentially the sense of "knowledge" previously introduced in 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.

4.2.3 Alan Turing: the $\mathbf{A}_T$-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 and extended 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.[2]

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.[3] 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

[2]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).

[3]Again, what I call a ULM_T is now most commonly referred to as a *Universal Turing Machine* (Minsky 1967; Lewis & Papadimitriou 1981).

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. They will be taken up again in due course. For the moment, however, I note simply that 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, these issues prove to be more or less irrelevant to von Neumann's work.

4.2.4 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 here 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 therefore also mandated von Neumann's analogous usage (in "universal constructor", or UCM in my terms). The typical operations of the two kinds of machine (computation and construction, respectively) are, of course, quite different. This is an important point, which I shall elaborate.

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.

4.2.4.1 Universal the First

The first connotation of "universal" in ULM, the one already described, and which is correctly preserved in von Neumann's analogy, refers simply to a relationship between the ULM and *all A-machines within its own A-system.* In my view this was the primary, if not the only, connotation which Turing had in mind when he first introduced the term "universal machine" . In any case, I suggest that this is the *only* connotation which von Neumann properly intended should carry over to the interpretation of UCM, as already described.[4]

4.2.4.2 Universal the Second

The second interpretation of "universal"—and the first which it would be erroneous to impute to the UCM—revolves around the idea that what makes a ULM "universal" is not *just* that there exists *some* relationship between it and some complete set of A-machines, but that there exists a very *particular* relationship—namely that of being able, when suitably programmed, to carry out the same A-computations. To put it another way, the "universality" of the ULM is seen to be *inseparably* bound up with the idea of "computation", so that it is not so much a matter of spanning a set of (A-)machines, but rather to be specifically about spanning a set of (A-)computations.

Now this is not an entirely *un*reasonable interpretation of "universal"—as long as we restrict attention to U*L*M's; because, in that case, it is entirely compatible with the original interpretation. However, in contrast to that original interpretation, the application of this second interpretation in the case of a U*C*M is deeply

[4]A further, fine, distinction *could* be made here between the idea of a ULM spanning *all* A-machines, and its spanning just those which can be regarded as realising some A-computation. This distinction does not arise in the A_T-system, because *all* A_T-machines *are* regarded as realising some A_T-computation. Fortunately (!) this is not a significant issue insofar as the analogy with the UCM is concerned, so I shall not pursue it further.

problematic and counterintuitive. If we try to force this interpretation, we come up with something vaguely like the following: given any (A-)computation, a UCM can, when suitably programmed, construct an a A-machine which could, in turn, carry out that (A-)computation.

At first sight, this is such an abstruse view of how the ULM and UCM might be related that one is inclined to say that it could not possibly arise. After all, von Neumann's whole point is to talk about automata which can construct automata *like* themselves; whereas, under the interpretation of the previous paragraph, the definition of a UCM would make no reference at all to its ability to construct automata "like itself" (i.e. which could, in their turn, also construct further automata "like" themselves), but would instead talk about the ability of a UCM to construct automata of a *different* (perhaps *very* different) kind—namely, "computing" automata.

Nonetheless, precisely this interpretation *has* been adopted in some of the literature, as we shall see. To explain how, and perhaps why, this arises, it is first useful to distinguish three variants on the idea, which differ in exactly how the "universal" set of "computations", which is to be spanned by the offspring of the UCM, is defined:

- In the simplest case, we assume that the A-system, in which the putative UCM exists, itself supports some definite notion of computation, which is to say it defines *some* set of A-computations. We then require only that the offspring of the UCM span this set. Specifically, we place no *a priori* constraints or requirements on what kind of thing should qualify as an A-computation.

- In the second case, we require that the set of A-computations of the A-system be such that, in some well defined sense, for every A_T-computation there must be at least one A-computation which is "equivalent". I shall omit any consideration of how such a relationship might be practically established. Given that it *can* be established, we then require that the offspring of the UCM span some set of A-computations which is "equivalent" to the set of A_T-computations (this may, or may not, be the complete set of all A-computations). On this interpretation, a UCM is related not to the

"general" notion of a ULM, but to the specific case of a ULM_T (i.e. a ULM in the A_T-system).

- Finally, we might require that the set of A-computations of the A-system be such that, in some well defined sense, for every "computation" of *any* sort, which can be effectively carried out at all, there must be some A-computation which is "equivalent". Again, I omit any consideration of how this relationship might be practically established. Given that it *can* be established, we then require that the offspring of the UCM span some set of A-computations which is "equivalent" to the set of all effective computations (and again, this may, or may not, be the complete set of all A-computations).

I refer to all three of these (sub-)interpretations of the "universal" in UCM as being "computational". In my view, of course, they are all three equally erroneous.

The first two of these computational interpretations of UCM could, in principle at least, be completely formalised in particular A-systems, so that the existence of a UCM in these (somewhat peculiar) senses would, at least, be a matter of fact, which might admit of proof or disproof.

However, the third computational interpretation relies on the informal notion of what constitutes an "effective computation", and will always be a matter of opinion or convention rather than fact; there is no possibility of the existence (or otherwise) of a UCM, in *this* sense, being decisively established for *any* A-system.

Having said that, Turing, in his original paper Turing (1936), argued (informally, of course) that the A_T-system already captures everything that could "reasonably" be regarded as an effective computation. As well as informal arguments to this effect, Turing showed that an equivalence could be established between the set of A_T-computations and the set of (A-)computations of an entirely different formalism proposed by Church. Similar equivalences have since been demonstrated with respect to a number of other independent formalisations, and the idea that the A_T-computations capture, in some sense, all possible computations, is now referred to as the *Church-Turing thesis* (e.g. Hofstadter 1979, Chapter XVII). Due to the necessarily informal nature of the claim, it is a *thesis*

not a *theorem*; nonetheless it is now widely regarded as being well founded (e.g. Minsky 1967, Chapter 5).

Now *if* the Church-Turing Thesis is accepted, then the third (computational) interpretation of UCM described above becomes exactly equivalent to the second. Indeed, one may say that the only reasonable basis for introducing the second computational interpretation at all is on the understanding that the Church-Turing thesis holds, because this implies that the A_T-computations provide an absolute benchmark of *all* kinds of computation. If this were *not* the case, then it would appear rather arbitrary to single out *this* set of computations for special significance relative to the notion of UCM.

More generally, it seems to me that it is *only* in the context of the Church-Turing Thesis that a strictly computational interpretation of the "universal" in UCM suggests itself at all. The point is that a ULM_T is (by definition) capable of carrying out all A_T-computations; and therefore, under the conditions of the Church-Turing Thesis, a ULM_T is, in fact, capable of carrying out all effective computations. We should perhaps say that a ULM_T is *doubly* universal: it is firstly universal with respect to all A_T-computations (which gave it its original title); but this then turns out (at least if the Church-Turing Thesis is accepted) to mean that it is universal with respect to the computations of *any* effective computing system whatsoever, not "just" those of the A_T-system. To make this completely clear, we should perhaps refer to a UULM, or U^2LM; but, since there is apparently no conflict between these two distinct attributions of universal (i.e. since the Church-Turing Thesis asserts that they are synonymous) it has become conventional not to bother to distinguish them; the single "U" in ULM_T (i.e. in "universal Turing machine") is, today, flexibly interpreted in either or both of these two senses, as the context may demand, without any further comment. I suggest that it is *only* because these two connotations of "universal" in ULM_T are not normally distinguished, that a strictly computational interpretation of "universal construction", or U*C*M, (i.e. any of the three such interpretations I have distinguished above) is typically entertained at all.

I stated that computational interpretation(s) of UCM have appeared in the literature. It is not always possible to isolate exactly which of the three identified sub-cases are intended, though this is not critical for my purposes, since, as

already noted, I consider them all to be mistaken. In any case, the most explicit (and, to the best of my knowledge, the earliest) advocate of a computational view of the UCM concept is E.F. Codd, and his proposal is quite precise, corresponding exactly to what I identified above as the second computational interpretation:

> The notion of construction universality which we are about to formalize demands of a space the existence of configurations with the ability to construct a rich enough set of computers such that with this set any Turing-computable partial function on a Turing domain can be computed in the space.
>
> Codd (1968, p. 13)

Codd's interpretation of UCM has been explicitly repeated by Herman (1973). Langton (1984) does not explicitly endorse Codd's interpretation as such, but does equate Codd's concept with von Neumann's, which I consider to be mistaken.

I should admit that it will turn out that the position, typified here by Codd, is not quite as perverse as I have painted it. Codd had special reasons for his particular approach,[5] and, even aside from these, it *will* ultimately prove useful to say something about the "computational" powers of A-constructors and/or their offspring.

However, my claim is that such powers should form no part of the essential *definition* of the UCM concept; in particular, they seem to me to be no part of von Neumann's *analogy* between the ULM and the UCM. While Codd's definition cannot, of course, be said to be "wrong", it is certainly *different*, in a substantive way, from von Neumann's; more seriously, we shall see that adopting such an interpretation would fatally undermine von Neumann's proposed solution to P_v. Since Codd does not say any of this, and since his work is otherwise explicitly based on that of von Neumann (Codd 1968, Introduction), his subsequent development is potentially misleading. This is all the more unfortunate as Codd *did* achieve certain significant new theoretical results.

To put this in a slightly different way, note that Turing and, equivalently, Church, proposed their thesis for a very definite reason. They were each attempting to solve the so-called *Entscheidungsproblem*, the *decision* problem of (meta-)mathematics, originally formulated by Hilbert.[6] The statement of this

[5] He was *inter alia* interested in the uses of "real" cellular automata as massively parallel *computers*.

[6] For a concise discussion see, for example, (Hodges 1983, pp. 91–94).

problem explicitly referred to the (informal) notion of a "definite method", or an "effective procedure" as it is now called; thus Turing's work could conceivably be regarded as a solution of this problem *only* if the Church-Turing thesis were accepted. The thesis was thus absolutely central and essential to Turing's analysis. Von Neumann's problem, on the other hand (at least in my formulation as P_v), makes *no* reference whatsoever to computation, "effective" or otherwise; so the Church-Turing thesis can have no *essential* rôle to play in its solution.

4.2.4.3 Universal the Third

I now come to the third (and final) distinct interpretation of "universal" (in UCM). This again involves the Church-Turing thesis, but in a way which is quite different from the strictly computational interpretations just outlined.

Roughly speaking, the Church-Turing thesis says that the computations of which A_T-machines are capable are universal with respect to *all* computational systems—regardless, for example, of their "material" structure. We could therefore attempt to, as it were, carry over this whole thesis, through von Neumann's analogy, to say something, not about *computational* systems in general, but *constructional* systems in general.

Now it is clear that von Neumann must indeed have had *something* at least vaguely of this nature in mind; for he hoped to establish the absence of paradox in the growth of complexity in the *biological* world, and this part of his argument can go through only if, in some sense, his results *transcend* the specific formalism or axiomatisation in which they are originally derived. On the other hand, the degree of generality actually required here is very weak. Von Neumann's only claim was that there is no *necessary contradiction* between the growth of complexity in the biological world, and the possibility of *some* strictly reductionist explanation of that world. This claim can be justified provided only that von Neumann can exhibit the possibility of such growth of complexity in *some* formalisation of automata theory: it is *not* required that this formalisation be particularly faithful or accurate as a representation of physical or biological reality.

More specifically: we shall see that von Neumann introduced the notion of a UCM as an element of his argument for the possibility of growth in automa-

ton complexity, but that, in considering how his results related to the biological world, von Neumann implicitly denied that UCM's *per se* play a rôle in biological organisms (von Neumann 1951, p. 318), thus leaving entirely open the question of whether "biological UCM's" (however they might be defined) are even possible, in principle.

Thus, von Neumann never attempted to formulate an explicit analog to the Church-Turing Thesis, incorporating the notion of construction (in place of computation); and insofar as he touched on the issue at all, it was in terms very much weaker than the Church-Turing Thesis. I therefore take the view that, although there is a strong and genuine analogy between von Neumann's work and Turing's, this has not been (and perhaps cannot be) extended to include any reasonable analog of the Church-Turing Thesis. To put it another way, whereas Turing claimed that the set of all A_T-machines (and thus any single UTM_T) was "universal" with regard to all effective computations, of *any* A-system, there is no analogous claim relative to the constructional powers of the set of all A-machines (or, equivalently, any single UCM) in any particular A-system (whether described by von Neumann or otherwise).

The point of this discussion is that the analogy between the UTM and UCM concepts is so strong that, until the issue is considered explicitly, one can be easily lulled into supposing that there *is* some obvious generalisation of the Church-Turing thesis; which would imply, in turn, that a UCM, in *any* "sufficiently powerful" A-system, captures something important about the powers of *all* automata, in *all* formal frameworks, and, by implication, about the powers of all "real" (physical) automata. It is important to emphasise that von Neumann himself never asserted, much less argued for, any such thesis; and that, for what it is worth, it seems unlikely (to me) that such a thesis could be defended. Conversely, to *assume* that some such thesis holds will be confusing at the very least, and also liable to lead to actual error in interpreting the implications of von Neumann's work.

Admittedly, as far as I am aware, no worker has ever *explicitly* argued for such a generalisation of the Church-Turing thesis—but there are some indications of its having been at least implicitly assumed.

Thus, Thatcher (1970, pp. 153, 186) makes passing reference to such a possibility, though he does not explore it in any detail. More substantively, while Tipler (1981; 1982) does not explicitly mention the Church-Turing thesis, he does interpret von Neumann's work as having extremely wide-ranging applicability, well outside anything actually mentioned by von Neumann himself. In brief, Tipler cites von Neumann as establishing that a "real", physical, UCM, which can construct *any* physical object or device whatsoever (given an appropriate description, sufficient raw materials, energy, and, presumably, time), can be built. It seems to me that such a claim must implicitly rely *inter alia* on something like a generalised Church-Turing Thesis; it is, in any case, directly contrary to von Neumann's comment, in discussing the general nature of his theory, that "Any result one might reach in this manner will depend quite essentially on how one has chosen to define the elementary parts" (von Neumann 1966a, p. 70).[7]

4.2.4.4 And So?

To conclude this discussion of "universal" construction: 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); but we can already state that 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.

[7] I claim, incidentally, that Tipler's interpretation of von Neumann's work can *separately* be severely criticised on a variety of other grounds. Some of these should subsequently become apparent; but to attempt a comprehensive critique of Tipler's work at this point would be a confusing distraction.

4.2.5 von Neumann's Solution

4.2.5.1 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)

As previously mentioned, Von Neumann's initial, informal, attempted solution to P_v was presented originally 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 his audience (Burks 1966d, 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 1966d, 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.

4.2.5.2 Some Notation

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

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, there will exist some unique $m_3 \in M_u$ 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 $m \in M_u$

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 a "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_u. 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. $u_0 \in M_u$), 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 $\forall\ m \in\ M_u$ 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:

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

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 $(u \oplus d(m))$ 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.[8]

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.[9] In fact, we do *not* (for the moment) require any

[8] 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.

[9] 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*.

one-to-one correspondence (for example) between the A-descriptors and A-tapes; which is to 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.

4.2.5.3 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.[10]

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:

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

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_u 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 $X = M_u$ as such, we can reasonably suppose that if M_u 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.

[10]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 powerful as a UCM!

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:

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

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.[11]

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 concept 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.

4.2.5.4 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

[11] 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 4.3.2.2 below.

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 (1969b, 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:

$$(u \oplus d(u)) \rightsquigarrow u$$

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

$$(u \oplus d(u)) \rightsquigarrow (u \oplus d(u))$$

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.

Before considering the correction which von Neumann found to overcome this, it is worth exploring the difficulty of a "direct" approach. This will demonstrate the claim, made earlier, that, although the correction ultimately proves to be minor, it is by no means a trivial matter to find it.

Let us denote by C_0 the set of A-constructors consisting of our basic UCM, u_0, composed with the A-descriptor of *any* A-machine $m \in M_u$.

$$C_0 = \{(u_0 \oplus d(m)) | m \in M_u\}$$

Our earlier, putative, A-reproducer corresponding to u_0 is one particular element of this set, namely $(u_0 \oplus d(u_0))$. We now see that this is, unfortunately not an A-reproducer after all. But, it might suffice for our argument if we could

guarantee, simply from the universal construction property of u_0,[12] that *somewhere* among the elements of C_0 there must always be at least *one* A-reproducer. This is to say, we might speculate (naïvely, as it will turn out), that even though u_0 composed with its own A-descriptor is not an A-reproducer, this set C_0 *will* contain at least one A-reproducer; which is to say at least one *fixed point* (under the action of $\rightsquigarrow$). This seems not altogether implausible because, after all, we regard C_0 as being rather large and diverse—recall that, for *every* A-machine $m \in M_u$, there is *some* element of C_0 which constructs it.

So: a "direct" approach to correcting the earlier error would then consist in establishing (from the property of universal construction) that every set of the form C_0 does include at least one A-reproducer. That such a direct approach *would* be naïve, at best, is shown by the following considerations.[13]

In attempting this direct approach, we are, in effect, trying to *directly* overcome the (apparent) paradox of self-reproduction, as originally formulated by von Neumann. Specifically, we can fairly easily accept the possibility of something like $(u \oplus d(u)) \rightsquigarrow u$, because it *does* involve a degradation in A-complexity; a UCM *without* any A-descriptor attached plainly *is* less A-complex, in some reasonable sense, than a UCM *with* an A-descriptor attached. So the reason that our original proposal for an A-reproducer *fails* to actually self-reproduce seems to be precisely an instance of degenerating (A-)complexity.

Let me try to make this even more explicit. The problem with $(u_0 \oplus d(u_0))$ is that it constructs just u_0 instead of $(u_0 \oplus d(u_0))$. Now $(u_0 \oplus d(u_0))$ *is* itself some A-machine in its own right—say $c \in C_0$; so if we want to construct c, perhaps we should program u_0 with $d(c)$ (instead of merely $d(u_0)$)? This *seems* like an improvement; at least now the offspring does have an A-tape attached. But, of course, we have only displaced, rather than eliminated, the problem. The parent A-machine is now $(u_0 \oplus d(c))$ instead of c itself (i.e. $(u_0 \oplus d(u_0))$, and, in turn, the second generation offspring (i.e. c's offspring) is not c either, but is simply

[12] Thus ensuring that a similar guarantee would then apply to *every* $u \in U$, as required by the core argument.

[13] It seems clear that von Neumann himself did consider (and reject) this naïve approach before hitting on his alternative approach (still to be discussed) which actually works. However, the only explicit discussion, of which I am aware, by von Neumann on this topic (von Neumann 1966b, p. 118) is quite cursory, and I shall try to fill out the arguments in rather more detail here.

u_0 with no A-tape again; we still have just further examples of degenerating A-complexity. We plainly cannot identify an A-reproducer by this procedure; nor, indeed, by any further iterations of it.

We may now begin to suspect that the paradox is a genuine one—at least in the restricted sense that even if self-reproduction is not paradoxical in general, it *is* paradoxical for all elements of C_0, i.e. all A-machines having the simple architecture $(u_0 \oplus d(m)))$. While, if true, this would be a rather negative conclusion, it *might* still represent progress (by eliminating things which will not work), and deserves some consideration for that reason.

In more detail, the argument *for* paradox here is roughly this: suppose firstly that some $c = (u_0 \oplus d(m))$ *is* self-reproducing. Then it seems that some "part" of the A-descriptor $d(m)$ must be taken up with describing u_0, with the "remainder" (presumably) describing the A-tape to be connected to u_0 in the offspring; but this latter A-tape is supposed to be precisely $d(m)$ again (on the assumption that c is indeed self-reproducing) and this means that a proper "part" of $d(m)$ must, in some sense, code for the whole of $d(m)$ itself. This certainly sounds like something dangerously close to paradox.

In fact, we can now perhaps see that the situation stops just short of any *necessary* paradox. It *may* indeed be the case that, for any certain *particular* "description language", no A-descriptor can contain a proper part which can serve *inter alia* to describe the A-descriptor as a whole—i.e. self-reproduction may actually be paradoxical for a specific set C_0 (relative to a specific UCM u_0—and thus also relative to all $u \in U$, sharing the same formal language); but there appears to be no valid argument showing that this must be so *in general* (i.e. for *all* UCM's, or all "possible" formal languages). Burks has made just this point, saying that:

> Prima facie it might seem that an automaton [A-machine] could not store a description of its own structure because, however many cells [A-parts] it had, storage of the description would require more than that number of cells ... This objection is of course not sound, because we may use indices, summation signs, and quantifiers in the description.
>
> Burks (1960, pp. 307–308)

However it should be clear from this that, while self-reproduction in some particular C_0 *may* not be actually paradoxical, this will be critically dependent

on the peculiarities of the description language processed by u_0. Indeed, it may seem that, even if one or more elements of (some) C_0 *are* A-reproducers, then this will be an essentially serendipitous or coincidental effect, almost certain to be disrupted by A-mutation; i.e. that even if we could exhibit some u_0 (and thus some set U) such that we could exhibit at least one A-reproducer "corresponding" to each $u \in U$ (which would seem like quite a tall order in the first place), the constraints imposed on the description language in order to achieve this may be such that the images of the A-reproducers *cannot* be kept "close" to each other in the space of A-tapes. That is to say that, *prima facie* at least, it seems that designing an encoding which guarantees the existence of A-reproducers at all may well conflict with the requirement that, under this encoding, the images of the A-reproducers must be "close" enough to each other to allow that they will be connected under some reasonable form of A-mutation.

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 (1969b, 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 $m \in M_u$, 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 UGM, and this is as follows. Given any A-machine $m \in M_u$, 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:

$$\begin{array}{rcccl} d' & : & M_u & \rightarrow & T \\ & & m & \mapsto & d'(m) \quad \text{s.t.} \\ & & & & (u_0 \oplus d'(m)) \rightsquigarrow (m \oplus d'(m)) \end{array}$$

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 $m \in M_u$ 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:

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

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.

4.2.5.5 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 (1966d, 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 A-descriptors of 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;[14] 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

[14]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".

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_u—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_u. 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_u 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_u, it is still an infinite set, spanning *essentially* the same range of A-complexity as M_u 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_u 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 attempted. 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_u. 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 $m \in M_u$ (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_u, 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_u 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_u) 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 associated with P_v; and since I will not pretend to be able to offer a really satisfactory answer, my treatment can be mercifully brief!

One possible answer—the only one (if any) which I think von Neumann himself could be said to have explicitly offered—is to say yes, M_u does span a sufficient range of A-complexity, *and this is self-evident*. This answer has, at least, the merit of an overwhelming simplicity. However, I think that it is possible to do better—though perhaps only very slightly.

I do not, of course, propose to formalise "A-complexity"; but following von Neumann's rough descriptions of the idea, and my own previous discussion in terms of equating it (more or less) with the notions of "knowledge" or "anticipatory systems" (McMullin 1992b, pp. 5–7), I propose[15] that A-complexity can be regarded as closely related to what Burks (1960) has called the *behavior*, or as I shall term it, the *A-behaviour*, of an automaton or A-machine.

A-behaviour *is* an essentially formal notion, and corresponds to the (real-time) specification of how an A-machine reacts to its environment. It is not, of course, a scalar quantity, and I shall not propose any measure of the A-complexity corresponding to particular A-behaviours.

The merit, for my purposes, of introducing the notion of A-behaviour, is that we can define a certain set of A-behaviours which, at least intuitively, captures our notion of what could, conceivably, be a "possible" A-behaviour (within any particular A-system); it constitutes, in short, a *universal* set, which I shall term B, of A-behaviours for that A-system. With this set B at our disposal, and without stipulating *how* A-complexity and A-behaviour might be related, we can say that if, for every A-behaviour $b \in B$, there is at least one $m \in M_u$ (or, even better, one $x \in X$) exhibiting this A-behaviour, then M_u (or X) must, in some sense, span all "possible" A-complexities, and therefore "must" meet the requirements of P_v (there is an essentially reductionist metaphysical assumption

[15] I shall *not* "argue" for this proposal; I shall merely tentatively adopt it as a basis for discussion.

underlying the interpretation of "A-complexity" here, but let it pass).

The universal set B of A-behaviours for a particular A-system may be roughly defined as follows. I assume that the definition of the A-system includes a specification of everything that might be regarded an an "environmental input" (A-sensor) or as an environmental output (A-effector) of any A-machine. I suppose that every A-machine has a fixed configuration of A-sensors and A-effectors (this is somewhat restrictive, but will serve my purposes here). An A-behaviour will then be completely defined by specifying a (finite) set of A-sensors and A-effectors, and a (hypothetical) finite state machine (see e.g. Minsky 1967, Part One) connecting these A-sensors and A-effectors together. Again, connecting A-sensors to A-effectors via a *finite* state machine represents something of a restriction, but it will serve my immediate purposes. The cartesian product of possible A-sensor/A-effector configurations by (compatible) finite state machines, will then yield the universal set of A-behaviours B for the particular A-system (and note carefully that because of the involvement of A-sensors and A-effectors, even if for no other reason, this definition *will* be tied to the particular A-system).

It should be clear that, if the set M_u of all possible A-machines in the A-system included elements realising the transition functions of arbitrary finite state machines (and assuming that these could be then "connected up" with arbitrary A-sensor/A-effector configurations), then we would have our desired result—M_u would span the range of all possible A-behaviours, and thus of all possible A-complexities, for that A-system.

One can actually envisage the possibility of formal A-systems of this sort (if, for example, our A-parts included the necessary elements to realise Burks' (1970c) *finite idealized automata*). But, it is certainly not the case that von Neumann's cellular model (for example) could meet this requirement (compare the remarks of Burks 1966a, p. 270).[16]

Let me then weaken the requirement somewhat. Let me require that, for every A-behaviour $b \in B$, there must be at least one $m \in M_u$ (or, better, $x \in X$) which can exhibit this A-behaviour *according to some sufficiently slowed down*

[16]More generally, it seems that no A-system which incorporates some principle of "local action" (i.e. the impossibility of "instantaneous" transmission of signals between arbitrarily "distant" A-parts) could meet such a strong requirement; but I shall not attempt to prove, or even to formalise, this claim.

time scale. That is, the A-behaviour can be realised if we consider the time-scale defining the A-behaviour (the clock rate of its finite state machine) to be scaled down to be as slow as we like compared to the actual (real-)time of the A-system.[17]

If this requirement, or criterion, for assessing the range of "A-complexity" spanned by M_u is accepted, then it can, for example, be met if M_u (or, better, X), includes at least one "universal" (in the Church-Turing thesis sense) *computing* machine (something with the computational power of a ULM_T) together with all its arbitrarily programmed variants (provided that this can be flexibly connected up with arbitrary A-sensor/A-effector configurations). *This* requirement can indeed be satisfied in the von Neumann cellular model; indeed, in Burks' completion of von Neumann's work, he specifically established that a single A-machine combining both a UGM and an (arbitrarily programmed) ULM_T (in effect) could be realised in this model. Burks has termed the latter a *universal computer-constructor* (Burks 1970b, p. xi).

As I said, I do not consider that this result really goes very much beyond a simple statement that the range of A-complexity spanned by M_u (or X) in von Neumann's cellular model (say) is "self-evidently" satisfactory for the solution of P_v. Indeed I feel that that (much simpler) answer has definite advantages. I introduce the alternative, somewhat convoluted, answer purely to show that here is one place where the solution of P_v *might* be said to directly depend on the "computational" properties of the A-machines. It provides a rationale—in my view the only valid one—for adding into the definition of "universal construction" something relating to ("universal") computation as such. It is the one point in the argument where it might even make sense to invoke the Church-Turing thesis as having some relevance. I fear I may be alone in this opinion, but that merely makes it the more important that I should state it as clearly as possible. In any case, I have now discharged the obligation I originally accepted in section 4.2.4 above to show that it would "... ultimately prove useful to say something about

[17] It is a very moot point whether, in so weakening the requirement one is not, perhaps, giving away rather *too* much; but I shall accept it without further discussion, simply to show where this can lead.

the 'computational' powers of A-constructors and/or their offspring". Of course, in doing so, I continue to reject entirely Codd's explicitly computational definition of "universal" construction.

4.2.6 Critique

The previous section presented von Neumann's *original* solution to P_v in some detail. I should now like to consider some elaborations—perhaps even improvements—to this solution. The gist of von Neumann's argument is that the existence of a single UGM, u_0, is (more or less) *sufficient* to allow P_v to be solved. My question is whether, or to what extent, we can weaken this condition—i.e. can we move closer toward a condition which is still sufficient, but also *necessary*. In doing this we shall get some glimpse of further important potentialities already implicit in von Neumann's solution.

As a starting point I take the requirement that the UGM u_0 should be a UCM in its own right. While this was indeed the case for the particular UGM exhibited by Burks, in his completion of von Neumann's work (Burks 1966a), it was not (or at least, not clearly) the case for von Neumann's own original formulation in the kinematic model, nor for his own outline for the cellular model (von Neumann 1966b, p. 119). The point is that the *only* property of the UGM which need actually be used (in solving P_v) is its ability to correctly process the *dashed* A-descriptors—i.e. to construct A-machines of the form $(m \oplus d'(m))$. Its ability to construct A-machines *not* having this structure (which is what additionally qualifies it as a UCM) is never actually used.

So: we can weaken the definition of the UGM, so that a UGM *need not* be a UCM (although it can be).

This is, of course a very minor improvement. Although we no longer require a UGM to be able to construct an arbitrary *isolated* A-machine, we still require that it be able to *embed* an arbitrary A-machine within its offspring. There is therefore no sense in which a UGM-but-not-UCM is likely to be *significantly* easier to realise, for example, than a UGM-and-also-UCM. So I introduce this merely as a clarification of the logical structure of the solution to P_v, rather than as anything of deep significance.

I now ask whether a *Universal* Genetic Machine, as such is truly required at all. And the answer, not surprisingly, will be that it "all depends". Let us consider a (non-universal) Genetic Machine, or GM, which I shall identify as g_0. The defining feature of g_0 is that it works *only* for some proper subset of the A-machines in M_u. That is, there exists some proper subset, $M_g \subset M_u$, such that for every A-machine $m \in M_g$ (and only for these) there exists a (dashed) A-descriptor $d'(m)$ which has the property that:

$$(g_0 \oplus d'(m)) \rightsquigarrow (m \oplus d'(m))$$

This obviously represents a weakening of the original requirement for a UGM; in fact, it essentially introduces a continuum along which this requirement can be weakened, depending on just how impoverished the set M_g becomes relative to M_u. What, if anything, can we say about how this will affect the solution schema for P_v? In particular, under what circumstances might the schema now fail?

Well, it seems clearly the case that we must have $g_0 \in M_g$, for otherwise there will not even exist the basic A-reproducer $(g_0 \oplus d'(g_0))$. So: it does not matter how extensive M_g *otherwise* is, it must at least contain g_0 itself.

More generally, let us interpret X in the same way relative to g_0 as it was originally interpreted relative to u_0 (of course, since g_0 and u_0 are different A-machines, this means that X is also now a more or less different set). That is, for every A-machine $x \in X$, composing this x with g_0 will not interfere with the latter's basic constructive processes. More concisely, we can still say that:

$$\begin{aligned} &\forall\, m \in M_g, \\ &\forall\, x \in X, \\ &\qquad\qquad ((g_0 \oplus x) \oplus d'(m)) \rightsquigarrow (m \oplus d'(m)) \end{aligned}$$

Corresponding then to the original set U of UGM's related to u_0, I shall denote the set of GM's related to g_0 by G:

$$G = \{(g_0 \oplus x) | x \in X\}$$

As in the case of u_0 and U, we stipulate that g_0 itself is also a member of G.

Now, as pointed out above, we had to require that $g_0 \in M_g$ to ensure that even a basic A-reproducer, incorporating g_0 would exist. We can now generalise this

as follows. Consider the set $G \cap M_g$. We are guaranteed that for every A-machine $g \in (G \cap M_g)$ (if any), there will exist a corresponding A-reproducer, namely:

$$(g \oplus d'(g)) \rightsquigarrow (g \oplus d'(g))$$

In effect then the set of A-machines $G \cap M_g$ completely characterises the set of A-reproducers which will be guaranteed to exist as a consequence of the existence of g_0 itself. So the question of whether any g_0 (which is *not* a fully fledged UGM) will suffice to solve P_v reduces to the question of whether this set $G \cap M_g$ still spans the required range of A-complexity, and whether we can still assume that the set of corresponding (dashed) A-descriptors will be connected under A-mutation. The latter is not entirely trivial: we would expect that, as the set $G \cap M_g$ is made smaller or more impoverished (by weakening the powers of g_0) then the corresponding set of A-descriptors may naturally become sparser in the space of A-tapes, and may therefore cease to be connected (even "approximately") under A-mutation.

We can identify two extremes here.

Suppose firstly that $G \cap M_g$ is essentially equal to our original set U (i.e. g_0 is "almost" a UGM). Stipulate that the "original" solution to P_v was accepted—i.e. we were satisfied that the original set X, and thus U, spanned a sufficient range of A-complexity, and the (dashed) A-descriptors corresponding to the elements of U were sufficiently close together (in A-tape space) to form a connected set under A-mutation. Then, assuming that the (dashed) description languages processed by u_0 and g_0 were essentially similar, g_0 would certainly suffice to solve P_v. Of course, under this assumption, the powers of u_0 have only been very slightly weakened to yield g_0; g_0 can do everything u_0 can do except (possibly) construct some A-machines in which are embedded A-machines *not* in the set X (where X is now being interpreted as essentially the same set relative to both g_0 and u_0). So, we may say that g_0 can indeed be something short of a fully "Universal" GM, and still be "equally" satisfactory in solving P_v. But the weakening represented by this seems quite minimal. Anyway, since X itself is extremely difficult to characterise it seems extremely unlikely that it would be any easier to design a GM with just this property than to design a full blown UGM.

The other extreme is represented by supposing that $G \cap M_g$ has only a *single*

element—g, say.[18] This suffices to establish the possibility of self-reproduction, though by an extraordinarily convoluted path! If we wish, we can identify it as the *necessary* and *sufficient* condition for what we may call "genetic" self-reproduction (it is not, of course, a necessary condition for self-reproduction *per se*: von Neumann's (1966a, p. 86) "growing crystals" would not satisfy this condition, for example). But it is still, in von Neumann's sense, a strictly *trivial* form of self-reproduction, despite its being called "genetic". By definition, the A-reproducer ($g \oplus d'(g)$ is not connected (by A-mutation) to *any* other A-reproducer (not, at least, any based on the same core GM, g_0), never mind being connected, directly or otherwise, to a set of A-reproducers spanning a "large" or "infinite" range of A-complexities. In terms of P_v, a GM g_0 giving rise to only this one A-reproducer would be of absolutely no interest whatsoever.

I note, in passing, that notwithstanding this, just such an ultimately impoverished GM[19] *has* been reported in the literature (Langton 1984; 1986). On the basis of my discussion, this kind of A-machine, in itself, can serve no purpose whatsoever relative to solving P_v; Langton, on the other hand, seems to imply that it can, but I suspect this to be another example of the continuing, damaging, influence of what I have already labelled the *von Neumann myth*. I shall return to this point in the next section.

Between the two extremes mentioned above for the content of the set $G \cap M_g$ (and thus, implicitly, for the power or A-complexity demanded of the core GM g_0) there remains a continuum. We may speculate that, in any particular A-system, it might be possible to identify a g_0 which is "significantly weaker" (in some sense) than a UGM, but yet is still powerful enough (in terms of the set $G \cap M_g$ it supports) that we would still regard it as providing, through von Neumann's schema, a satisfactory solution to P_v. This would ultimately depend on informal judgements as to the range of A-complexity spanned by the set $G \cap M_g$, and as to whether the corresponding set of (dashed) A-descriptors is still sufficiently well connected (under A-mutation in A-tape space). Given that these judgments

[18]Note that this condition does *not* imply that either G or M_g are, in any sense, "small" sets; the underlying, or core, GM g_0 could still be, in this sense, very "powerful". Nonetheless, we could reasonably expect that it would be easier to design this GM than a full blown UGM. The question, of course, is whether there might be any benefit in so doing!

[19]Perhaps I should say *pen*ultimately impoverished: an *ultimately* impoverished GM would have $G \cap M_g = \emptyset$!

are informal, there can be no question here of providing any clearcut, definitive, criterion which would be both sufficient and *necessary* for the successful application of von Neumann's solution schema within any particular A-system. The most that we can say in general seems to be this: if, in a particular A-system, the existence of a UGM would indeed suffice for the solution of P_v (and even this judgment will always involve a degree of informality, as we have discussed), then it seems that something less powerful than a UGM should still suffice; but that the question "How much less powerful?" will not admit of any sharp answer.

Insofar as this analysis yields any substantive result it is simply that there is nothing decisive about the notion of "Universal" (genetic) construction as such; quite aside from the fact that the significance of this "Universal" will vary from A-system to A-system, it does not even play a unique or distinctive rôle in the application of the solution schema to a particular A-system. In fact, with many "reasonable" axiomatisations of the notion of A-machine, "universal construction" (and thus UGM's) may be literally *impossible*. This follows from Moore's (Moore 1970) so-called *Garden-of-Eden* theorem. This theorem applies, indeed, to von Neumann's cellular model, although this point is disguised by von Neumann choice of a somewhat restrictive "universal" set of A-machines (i.e. a set which excludes certain entities which might *intuitively* be regarded as perfectly reasonable A-machines)—see Burks' discussion of this (Burks 1970d, pp. 43–45). In terms of P_v there is *no* especially unique or distinguished, or "intuitively reasonable", notion of "universality".[20]

I draw this point out because we have seen that, simply by referring to "universal" construction at all, von Neumann opened up a large field of potential confusion. Granted, von Neumann evidently wanted to make clear an intellectual debt to Turing's original "universal (computing) machine". But with hindsight we can now see, perhaps, that the debt is really not so great as all that: whatever analogy existed between the ULM and UCM, it became significantly more strained or remote when referred to the UGM; and, arguably, becomes positively misleading when finally referred, as in this section, to a merely "sufficiently powerful" GM. In deference to von Neumann's example I have, in previous sections,

[20]That is: comparable to, say, the notion of "universal (effective) computation" associated with the *Entscheidungsproblem* and the Church-Turing thesis.

resolutely followed the essential sequence of his original solution to P_v, including all the distracting discussion of "universality"; but, having now done that, I venture to suggest that the solution could be made significantly more transparent by *starting* simply with the notion of a ("sufficiently powerful") GM, rather than, by tortuous paths, *ending* there.

The discussion thus far has been conducted entirely within the scope of von Neumann's original schema; it has consisted in little more than elaborating somewhat more precisely the conditions under which the schema becomes applicable (though, of course, that is a useful enough exercise in itself). But I should now like to point out that von Neumann's schema can be substantially generalised (at least in the abstract, or "in principle"), and that doing so can yield some significant benefits.

Consider again, then, a basic GM, g_0. This GM will give rise to a more or less diverse set of A-reproducers of the form $(g \oplus d'(g))$ as already discussed at length. For the moment I make no assumption as to how large or small this set may be; g_0 could, in one limit, be a full blown UCM, and the set would accordingly be expected to be very large; or, in the other limit, g_0 could be a very weak GM, yielding only a handful of A-reproducers. Whatever this set is, that completely determines whether or not that g_0 will be able to deliver a solution to P_v according to the von Neumann schema; and if the answer is "not" then that g_0 is essentially of no further interest.

Now this set of A-reproducers anchored on a single g_0 have precisely this in common: they process the same formal language for describing A-machines. In biological terms we may say that this set incorporates a fixed, or *absolute* mapping between genotype (A-descriptor) and phenotype (A-reproducer). Thus, in committing ourselves (following von Neumann) to solving P_v purely within the resources of a single such set, we are also committing ourselves to the equivalent of what I have elsewhere called *Genetic Absolutism* (McMullin 1992c, Section 5.3), within the analysis of our formal or artificial A-system.[21] I should note that, in that paper, I argue at length against the idea of Genetic Absolutism; but not in the sense that it is "bad" in itself—it just is not a tenable theory of biological

[21]Note carefully that this is strictly a limitation of the way *we choose to analyse* an A-system; it need not, and generally will not, reflect an inherent limitation of an A-system *in itself.*

evolution. Now von Neumann is not yet trying to capture all the complications of biological evolution: he is merely trying to establish that some key features, at least, can be recreated in a formal, or artificial, A-system. If this can be done within what is, in effect, a framework of Genetic Absolutism, *and if there is some advantage to doing this in that particular way,* then the fact that it is still "unbiological" (in this specific respect) should not be held too severely against it. Indeed, we shall recognise much more severe discrepancies than this when, in due course, we examine the new problem situation created by the solution of P_v.

Now, as it happens, adopting Genetic Absolutism *does* have a significant advantage for von Neumann. Working within such a framework it *is* necessary to exhibit one core GM, g_0; and it *is* necessary to establish that this is sufficiently powerful to satisfy the informal requirements of P_v; and it *is* finally necessary to show that, based on the formal language processed by g_0, there is a reasonable likelihood that most, if not all, of the corresponding A-reproducers will be directly or indirectly connected under A-mutation. But if all this can be done, then P_v can, indeed be solved. What would be the alternative if Genetic Absolutism were not adopted?

Well, the alternative to Genetic Absolutism is *Genetic Relativism* (McMullin 1992c, Section 5.4), which envisages that the mapping between genotype (A-descriptor) and phenotype (A-reproducer) is *not* fixed or absolute but may vary from one organism (A-reproducer) to another. If we tackle P_v in a framework of Genetic Relativism, we do *not* restrict attention to a single GM, giving rise to an "homogenous" set of A-reproducers, all sharing the same description language. Instead we introduce the possibility of having many *different* core GMs—g_0^1, g_0^2 etc. Each of these will process a more or less *different* description language, and will thus give rise to its own unique set of related A-reproducers. We still establish that most if not all A-reproducers in each such set are connected under A-mutation; but, *in addition,* we try to show that there are at least some (A-)mutational connections between the *different* such sets. The latter is, of course, a much more difficult task, because the A-mutations in question are now associated with changes in the very languages used to decode the A-tapes. But, if such connections can be established, then, for the purposes of solving P_v we are not restricted to considering the range of A-complexities of any *single* set of

A-reproducers, but can include the union of the sets.

Now clearly, in terms simply of solving P_v, Genetic Relativism introduces severe complications which are not necessary, or even strictly useful. For now we have to exhibit not one, but multiple core GMs, processing not one, but multiple description languages; and we have to characterise the range of A-complexity, and A-mutational connectivity, of not one but multiple sets of A-reproducers; and finally, we still have to establish the existence of A-mutational links *between* these different sets of A-reproducers. The only benefit of any sort in this approach seems to be that maybe—just maybe—the distinct GMs can be, individually, significantly simpler or less powerful than the single GM required under Genetic Absolutism; but it seems quite unlikely that this could outweigh the additional complications.

Let me say then that I actually accept all this: that for the solution of P_v as stated, adopting the framework of Genetic Absolutism seems to be quite the simplest and most efficacious approach, and I endorse it as such. Nonetheless, I think it worthwhile to point out the *possibility* of working in the alternative framework of Genetic Relativism for several distinct reasons.

Firstly, it would be easy, otherwise, to mistake what is merely a pragmatic preference for using Genetic Absolutism in solving P_v with the minimum of effort, for a claim that Genetic Absolutism is, in some sense, *necessary* for the solution of P_v. It is not. More generally, our chosen problem, P_v, is *only* concerned with what may be possible, or sufficient—not what is necessary.

A second closely related point is this: *prima facie,* our solution based on Genetic Absolutism may seem to imply that a *universal* GM (or, at least, something not far short of that) is a pre-requisite to *any* evolutionary growth of A-complexity. It is not. Indeed, we may say that, if such an implication *were* present, we should probably have to regard our solution as defective, for it would entirely beg the question of how such a relatively A-complex entity as a UGM (or something fairly close to it) could arise in the first place. Conversely, once we recognise the *possibility* of evolution within the framework of Genetic Relativism, we can at least see how such prior elaboration of the powers of the GM(s) could occur "in principle"; this insight remains valid, at least as a coherent conjecture, even if we have not demonstrated it in operation. It precisely underlies the remark

already made that the advantage of Genetic Relativism in relation to the solution of P_v (insofar as there is one at all) is that it may permit us to work, initially at least, with significantly more primitive GM's as the bases of our A-reproducers.

Thirdly, Genetic Absolutism views all the A-reproducers under investigation as connected by a *single* "genetic network" of A-mutational changes. This is sufficient to solve P_v, as stated, which called only for exhibiting the *possibility* of A-mutational growth of A-complexity. In practice, however, we are interested in this as a basis for a *Darwinian* growth of A-complexity. Roughly speaking, this can only occur, if at all, along paths in the genetic network which lead "uphill" in terms of "fitness" (S-value). If the genetic network is fixed then this *may* impose severe limits on the practical paths of Darwinian evolution (and thus on the practical growth of A-complexity). Again, once we recognise the *possibility* of evolution within a framework of Genetic Relativism—which offers the possibility, in effect, of changing, or jumping between, *different* genetic networks—the *practical* possibilities for the (Darwinian) growth of A-complexity are evidently greatly increased.

This last point represents a quite different reason for favouring the framework (or perhaps we may now say "research programme") of Genetic Relativism, and it is independent of the "power" of GM's. In particular, even if we can exhibit a single full blown UGM, which yields an A-mutationally connected set of A-reproducers spanning (virtually) every possible A-behaviour supported in the A-system, there could still be advantages, from the point of view of supporting Darwinian evolution, in identifying alternative (U)GM's, defining alternative genetic networks (viewed now as evolutionarily accessible pathways through the space of possible A-behaviours).

Indeed, this need not be all that difficult to do: it provides another (in my view, much more compelling) reason to consider combining a basic (U)GM with a ULM_T (or something of similar computational powers): the latter is arranged so that it "pre-processes" the A-descriptor in some (Turing computable) fashion. The program of the ULM_T could then effectively encode a space of alternative description languages (subject to the primitive constructional abilities of the original (U)GM); with moderately careful design, it should be possible to open up an essentially infinite set of (U)GM's, which are themselves connected under A-

mutation (of the program for the embedded ULM_T—another A-tape of some sort), thus permitting a multitude of *different* genetic networks for potential exploitation by a Darwinian evolutionary process. This should greatly enhance the possibilities for Darwinian evolution of *any* sort, and thus, in turn, for evolution involving the growth of A-complexity.[22] This idea seems to have been anticipated by Codd:

> A further special case of interest is that in which both a universal computer and a universal constructor (*sic*) exist and the set of all tapes required by the universal constructor is included in the Turing domain T. For in this case it is possible to present in coded form the specifications of configurations to be constructed and have the universal computer decode these specifications ... Then the universal constructor can implement the decoded specifications.
>
> Codd (1968, pp. 13–14)

While Codd did not elaborate on *why* such flexibility in "coding" should be of any special interest, it seems plausible that he had in mind precisely the possibility of opening up alternative genetic networks.

I close this critique with two final remarks relating to Genetic Relativism.

Firstly, von Neumann himself seems to have discounted even the *possibility* of Genetic Relativism being applicable to his models. In his discussion of different kinds of (A-)mutations, he stated explicitly that A-mutations affecting that part of an A-descriptor coding for the core part of the A-reproducer (i.e. coding for g_0 in the terms used above) would result in the production of "sterile" offspring (von Neumann 1966a, p. 86): the implication is that this would *always* be the outcome of such A-mutations. I suggest that such a claim is too strong, in general. My view is that, on von Neumann's model, it is probably fair to say that such A-mutations would *almost* always yield sterile offspring; but that depending on the detailed design of the GM, and the nature of the particular A-mutation, there *might* be exceptional cases where the offspring would still be an A-reproducer, but containing an altered core GM.

Secondly, when tackling P_v within the framework of Genetic Absolutism, it was *necessary* to assume a degree of compositionality in the description language,

[22] It should be clear that this proposal is closely related to the more general suggestion already presented in Chapter 3, section 3.8.2, that the *efficient* growth of knowledge, via UVSR, will necessarily rely on the elaboration of a loosely hierarchical structure of variational processes.

to assure that there would exist a range of A-mutations *not* affecting the core GM in an A-reproducer; without this assumption it would be difficult, if not impossible, to argue that the set of A-reproducers anchored on this single core GM would be connected under A-mutation. This compositionality assumption is more or less equivalent to the biological hypothesis of *Genetic Atomism*, which holds that genomes may be systematically decomposed into distinct *genes* which, individually, have absolute effects on phenotypic characteristics (see McMullin 1992c, p. 11; Dawkins 1989b, p. 271). This again represents a divergence between von Neumann's pragmatically convenient solution schema for P_v, and the realities of the biological world (where any simple Genetic Atomism is quite untenable). I conjecture therefore that, should we wish to move away from a strict Genetic Absolutism in our formal or artificial systems we might well find it useful, if not essential, to abandon simple compositionality in our descriptive language(s) (i.e. Genetic Atomism) also. This, in turn, would ultimately lead away from A-reproducer architectures in which there is any simple or neat division between the core GM and the rest of the A-machine (though there might still be a fairly strict separation of the A-descriptor—i.e. a genotype/phenotype division).

4.2.7 The Von Neumann Myth

Having now presented and criticised, in some detail, what I have identified as von Neumann's solution to von Neumann's problem, I must discuss, once again, whether this really was the problem John von Neumann sought to solve. In one sense, of course, this is of no importance; provided P_v is admitted as an interesting and difficult problem, relevant to the interests of this Thesis, and provided that von Neumann did, indeed, solve it, then it hardly matters whether, as a matter of historical fact, von Neumann himself saw his work in precisely this way. But, in another sense, the question is very important; my stated reason for re-presenting von Neumann's work at length, and in detail, here was the claim that its significance has not been properly recognised, and that this has meant that his research programme (which is essentially now also my research programme) has foundered. This claim needs at least some further discussion and support.

Briefly, I conjecture that there exists what I may call a *von Neumann myth*: namely that, in his work on the Theory of Automata, von Neumann was concerned with some "problem" of self-reproduction *as such*, and/or that von Neumann proposed that universal computational abilities could provide a criterion of demarcation between "trivial" and "non-trivial" instances of self-reproduction. I admit, of course that von Neumann was concerned with *some* problem of self-reproduction; but in my view it was not self-reproduction as such, but self-reproduction *as a route to the spontaneous growth of complexity* (particularly via Darwinian evolution) that interested him; and that even though he was *also* immensely interested in the theory and practice of computing automata, the "computational abilities" *as such* (i.e. as opposed to the implications of such abilities for complex behaviour and/or evolutionary potential) associated with his self-reproducing automata were a matter of almost negligible importance.

Following on this, my task is twofold. First to back up my assertion that something like a von Neumann myth actually exists. And secondly to reiterate why I consider the position(s) identified with the myth to be untenable. I shall take these in reverse order because I think von Neumann himself was largely, if inadvertently, responsible for the origin of the myth; I shall therefore consider those elements of von Neumann's writings which might *seem* to give rise to the myth, and show how, in my view, they cannot be seriously upheld. Then I shall show how, notwithstanding this, the von Neumann myth has indeed been formulated and propagated.

Consider von Neumann's first published presentation of his ideas on a generalised theory of automata, taken from his Hixon symposium lecture in 1948. We may find there the following seemingly clearcut statement:

> The problem of self-reproduction can then be stated like this: Can one build an aggregate out of such [i.e. kinematic] elements in such a manner that if it is put into a reservoir, in which there float all these elements in large numbers, it will then begin to construct other aggregates, each of which will at the end turn out to be another automaton exactly like the original?
>
> von Neumann (1951, p. 315)

Out of context this certainly suggests that von Neumann's problem was *self-reproduction,* pure and simple. But, despite the somewhat unfortunate phrasing

and emphasis here, there *is* a context, which must not be ignored. Just two pages earlier von Neumann introduced the motivation for his work, at some length, as being the apparent paradox presented by the ability of biological organisms to maintain their complexity in self-reproduction, and for that complexity to increase "over long periods of evolution" (p. 312). Furthermore, even in this earliest paper, von Neumann went on, after explaining his scheme for how self-reproduction could be based on a UGM, to point out that this scheme had "some further attractive sides, into which I shall not go at this time at any length" (p. 317); while this further discussion was, indeed, brief, he did point out quite explicitly that his particular scheme of self-reproduction "can exhibit certain typical traits which appear in connection with mutation, lethally as a rule, but with a possibility of continuing reproduction with a modification of traits" (pp. 317–318). Whether von Neumann was originally led to his particular scheme by the need to support these things, or whether he was merely "sleepwalking"[23] is not really at issue here. The point is that he clearly recognised that his scheme offered a solution of a very difficult problem, namely what I have designated P_v, and he *did* say this, even if, with hindsight, we might wish he had been a little more explicit.

Similar remarks can be made about the Illinois lecture (von Neumann 1966a, Fifth Lecture), which I have already quoted several times in previous sections. Again, he introduced his problem as being the apparent paradox of the growth of complexity in the biological world. Again, a significant part of his discussion was *then* devoted to the "problem of self-reproduction", in the sense of establishing that self-reproduction could indeed be based on a UGM. But again, crucially, he concluded by discussing how this scheme supported mutational change, while still retaining the self-reproductive ability. Indeed, as previously noted, he even went to far as to explicitly cite "the ability to undergo inheritable mutations" as a criterion of demarcation between "trivial" and "non-trivial" reproduction. In my statement of P_v I elaborated this slightly by making explicit a requirement that such inheritable (A-)mutations connect up a set of A-reproducers of diverse A-complexities; but, taken in context, that was clearly already implicit in von Neumann's treatment.

[23]This evocative term seems to have originated with Koestler (1959).

Finally we come to what was to be von Neumann's *magnum opus* in the field, his unfinished manuscript *The Theory of Automata: Construction, Reproduction, Homogeneity*. Von Neumann here concisely outlined (on what is now just the second page of the published version, von Neumann 1966b) the complete set of 5 questions (labelled A–E) which he proposed to answer—or, in my terms, the problems he proposed to solve.

Von Neumann's question (A) was, admittedly, concerned with the computational powers of automata. But this was natural since computers were then by far the largest and most complex artificial automata which had yet been built; and, furthermore, von Neumann intended to introduce his UCM (and, subsequently, the UGM) by analogy with Turing's "universal (computing) machine", so that discussing computing automata would represent an essential preamble. Again admittedly, when it came time to answer the questions von Neumann did take care to ensure that what we might call "universal computational processes" could be realised in his cellular model. But I deny that any of this had any deep significance. Certainly, some kind of general purpose signal processing or computational abilities would be necessary if the A-machines were to be said to span a reasonable range of A-complexities; I have drawn this point out in detail myself. Such abilities would also be of direct assistance in actually designing the UGM. But, this is all really incidental to the central argument. I think it is much more significant to note that none of the remaining 4 questions, which represent the real substance of von Neumann's programme, made any reference to computation as such.

Von Neumann's questions (B) and (C) were concerned with the question of construction universality—specifically whether this could be demonstrated in some model system(s). In themselves, these questions were not explicitly motivated.

Question (D) introduced (at last) the "problem" of self-reproduction—but in a very special form. Von Neumann explicitly referred to this question as a "narrowing" of his question (C) relating to universal construction; this was, at the very least, a broad hint that he was not interested in self-reproduction *per se* but in self-reproduction which was built upon universal construction, though as yet there is no indication of why this should be of special interest. He implicitly

reinforced this interpretation by asking for self-reproducing automata which could "perform further tasks", such as constructing "other, prescribed, automata".

Finally, von Neumann's question (E) took up the question of evolution, and asked in particular whether automata "complexity" and/or "efficiency" can "progress".

My own view is that the only coherent or motivated way of viewing this programme of von Neumann's is to read it in reverse: starting with his question (E) of whether or how (Darwinian) growth of complexity can even be possible, and seeing all the other questions as merely subproblems, or intermediate goals, on the way to solving this fundamental problem. This interpretation makes sense on the assumption that von Neumann had already worked out, in outline at least, his solution schema for this fundamental problem; but of course, we *know* this to be the case because von Neumann had already presented the outline solution in the Hixon and Illinois lectures.

Now I should admit that there is one sense in which it seems von Neumann may have been genuinely concerned with what we might call a problem of self-reproduction "in itself". He conjectured (without elaboration) that self-reproduction based on direct self-inspection may be impossible, since otherwise "one would probably get involved in logical antimonies of the Richard type" (von Neumann 1966b, p. 122). His architecture based on the use of A-descriptors does indeed "solve" (or at least avoid) this problem. However, it now seems clear that von Neumann had here identified an entirely spurious *pseudo*-problem; Laing (1977) exhibited an early counter-example, showing how reproduction by self-inspection is, in fact, perfectly possible, without paradox.[24] Thus, in the one place where it seems that von Neumann could fairly be considered as tackling a "problem" of self-reproduction as such it now seems that he was actually mistaken.

That completes my argument for how von Neumann himself *might* have given rise to the von Neumann myth—and also for how the position(s) identified with the myth cannot be upheld. More particularly, I have not been able to identify anywhere where von Neumann discussed self-reproduction *except* in the con-

[24]Or at least, with no more paradox that von Neumann's own A-reproducers; cf. Rosen (1959).

text of an evolutionary growth of complexity; nor have I found anywhere where von Neumann proposed or adopted computational ability as criterial for "non-trivial" self-reproduction. He *did* adopt heritable, viable, mutation as criterial in this sense; and he *did* show that universal construction (in the form of a UGM) provides at least one way of achieving this (though not, of course, that it is the only way). In short, it seems to me highly unlikely that von Neumann could have *intended* to promote the views I have identified as the "von Neumann myth".

The *only* discrepancy of substance (of which I am aware) between what I have called P_v and the problem von Neumann described himself as being concerned with, is that von Neumann, in one brief note, considered how his A-reproducers might support *Lamarckian*, as opposed to Darwinian, evolutionary change (von Neumann 1966b, p. 131).[25] I omitted this from my formulation of P_v because Lamarckism is not, in itself, a satisfactory *biological* theory of the growth of organismic complexity (McMullin 1992b, Section 4.4). In any case, Lamarckism is not an element of my alleged von Neumann myth.

The remaining task here is to establish that a von Neumann myth does actually exist (and indeed persists).

I start with Kemeny's article (Kemeny 1955), based on von Neumann's Vanuxem lectures delivered at Princeton University in early 1953. I think we may say that the seed of the myth is already present here. For although Kemeny does refer to the question of realising an artificial evolutionary process, he does so only at the very end of the article, almost as an afterthought, and with no discussion of how von Neumann's *specific* scheme of self-reproduction addresses precisely this problem. There is also no discussion of the apparent paradox of the growth of complexity in the biological world. On the contrary, in fact, the thrust of the article seems to be to identify artificial self-reproduction in itself (no matter how realised) as the problem—and to then present von Neumann's work as a solution. Which is to say, a form of the myth.

However, the most telling source for the von Neumann myth, it seems to me, is A.W. Burks.

I should say, in advance, that I have the greatest respect for Burks; and that,

[25]Von Neumann does not use the term "Lamarckian evolution" as such, but that is effectively what he describes in the second paragraph of his section 1.8.

further, I owe him a considerable debt, for without his work, von Neumann's original manuscripts, upon which I have drawn very heavily, might never have been published; nor, perhaps, might the outstanding collection of seminal works in this field, *Essays in Cellular Automata* (Burks 1970a), have ever been brought together in one volume.

However, despite all this, I wish to suggest that Burks made a mistake. I conjecture that he did (perhaps still does) erroneously subscribe to the von Neumann myth; that this contaminated his work on the von Neumann's manuscripts which he (Burks) edited and completed; and that, as a result of the apparent authority of Burks' remarks, the myth has been indefinitely propagated.

The volume of von Neumann's work, edited and completed by Burks, collected together the manuscripts associated with von Neumann's Illinois lectures and his unfinished manuscript on automata theory, and was published as (Burks 1966d); a taste of the myth appears already in the title Burks chose for this collection: *Theory of Self-reproducing Automata*. Burks justifies this choice by referring repeatedly to von Neumann's work as being concerned with the problem of "self-reproduction", and describes the manuscript of Part II (von Neumann 1966b) exclusively as treating "the logical design of a self-reproducing cellular automaton (*sic*)" (Burks 1966c, p. xvi). On the same page, and without qualification, Burks makes the extraordinary remark that "self-reproduction *requires* an automaton of considerable complexity" (emphasis added). I call this extraordinary because, on my interpretation, the point of von Neumann's work was almost precisely opposite to this remark: far from showing that complexity was "required" for self-reproduction, von Neumann sought to establish how self-reproduction might still be possible *despite* arbitrarily high complexity.

However, I shall not attempt to identify every point at which Burks might be said to have supported, directly or indirectly, the von Neumann myth. It will suffice to identify what seem the most decisive examples.

Burks claimed that the central question addressed by von Neumann, particularly in (von Neumann 1966b), was "What kind of logical organization is sufficient for an automaton to be able to reproduce itself?" (Burks 1966b, p. 19). Even taking account of the full context, I cannot find any way of interpreting this claim other than as a statement that von Neumann's problem was some problem of

self-reproduction *per se*; which is to say, a statement of the von Neumann myth.

In completing his editorial work on (von Neumann 1966b) Burks added a final chapter (Burks 1966a) which included a *Summary of the present work* (Section 5.3.1). In this, Burks reviewed the five questions (A) through (E) which von Neumann originally started out with. With regard to question (E)—which I have argued provided the very essence and motivation of the entire manuscript—Burks says only that "Von Neumann made a few remarks relevant to evolution ... but never returned to the topic" (p. 287); the rest of Burks' summary then completely ignores this question. This is all the more striking when we contrast it with an earlier parenthetical remark by Burks' that "Whenever he [von Neumann] discussed self-reproduction, he mentioned mutations" (Burks 1966d, p. 99).

Burks went on, in his summary, to "reformulate" von Neumann's remaining questions in the context of von Neumann's particular cellular model, "at the same time modifying them somewhat" (Burks 1966a, p. 292). One of these modifications affects von Neumann's original question (D); where von Neumann had asked for a self-reproducing automaton which could do "other tasks", such as constructing "other, prescribed, automata", Burks now altered this to call instead for a self-reproducing automaton which can also "perform the computations of a universal Turing machine". I believe this may be the first occasion on which this *specific* idea was proposed—and, as far as one can tell, it was *not* proposed by von Neumann. Burks offered no explanation of the change at this point, but we see here another element of the von Neumann myth being born.

Moving on to (Burks 1970b, p. xv) we find a restatement of the earlier claim that von Neumann was seeking to answer the question "What kind of logical organization is sufficient for an automaton to be able to reproduce itself?" But now, at least, Burks points out that the question "admits to trivial versions as well as interesting ones"; he states that von Neumann had the "familiar natural phenomenon of self-reproduction in mind", with which I agree; but he then goes on to say that von Neumann "wished to abstract from the natural self-reproduction problem its logical form" which I consider to be obscure, at best. Burks does *not* mention here von Neumann's own formulation that the possibility of heritable, viable, mutation distinguishes the non-trivial form of the problem.

Turning finally to (Burks 1970d), we find first (p. 3) the now familiar claim

that von Neumann was concerned with the problem of a sufficient "logical organization" for self-reproduction. But, much more importantly, after a detailed discussion of the design of a UGM, and of an A-reproducer based upon it, in von Neumann's cellular space, comes this passage:

> This result is obviously substantial, but to express its real force we must formulate it in such a way that it cannot be trivialized. Consider, for example, a two-state cellular system whose transition function takes a cell into state "one" when any of its neighbors is in state "one". Define an automaton to be any area, even a single cell. A cell in state "one" then "reproduces itself" trivially in its neighboring cells. *Clearly what is needed is a requirement that the self-reproducing automaton have some minimal complexity.* This requirement can be formulated in a number of ways. We will do it by requiring that the self-reproducing automaton also be a [universal?] Turing machine.
>
> Burks (1970d, p. 49, emphasis added)

Here we have the von Neumann myth in its purest form. To be fair to Burks, he does not explicitly ascribe this position to von Neumann; but from the context, such an ascription would seem to be implied. The irony, again, is that von Neumann did address precisely the issue Burks raises here, when he spoke of the triviality of reproduction in "growing crystals" (von Neumann 1966a, p. 86); but, of course, von Neumann's resolution was nothing to do with computation. Instead, he identified heritable, viable, mutation as the critical criterion, which, in turn flagged his problem as my P_v, and *not* as self-reproduction *per se* (not even its "logical organization").

It seems to me that Burks' argument, on the other hand, can be understood only by firstly assuming, or *demanding* we might say, that von Neumann *was* trying to solve some "problem" of self-reproduction, and indeed that he did solve it; but then noticing that this is a pseudo-problem, admitting of trivial solutions; and finally trying to find some way of immunising von Neumann's obvious "success" from this criticism. There is, of course, a germ of truth in this view—my own analysis of von Neumann's work was arrived at in roughly this way. But, on my view, the correct resolution is not a direct requirement to embed some minimal "complexity" represented by a Universal Turing machine (say); this idea simply does not work because one can easily formulate a cellular space in which trivial (crystal-like) self-reproduction is still possible even for A-machines incorporating Turing machines (universal or otherwise). This is essentially the force

of a later paper by Herman (1973);[26] Herman concludes explicitly that:

> What the result does show is that the existence of a self-reproducing universal computer-constructor in itself is not relevant to the problem of biological and machine self-reproduction. Hence, there is a need for new mathematical conditions to insure non-trivial self-reproduction.
>
> Herman (1973, p. 62)

While, of course, agreeing with the essence of this, I disagree literally with the last sentence, which I consider to illustrate only the lingering after-effect of the von Neumann myth (apparently inherited by Herman, through Codd, from Burks). Perhaps we do need conditions to insure non-trivial self-reproduction, though I personally prefer to say that we need to reorient ourselves as to the problem we are tackling—and recognise that it is *not* helpful to describe it as a problem of "self-reproduction". But, in any case, we do *not* need "mathematical" (or formal) conditions. Not yet, at least. For we are not yet ready, by any means, to formalise "A-complexity"; and *that* (not "self-reproduction") is the point at issue. And, of course, as I have already repeated several times, von Neumann himself had already provided a perfectly serviceable *informal* condition, in the form of heritable, viable, mutations, so Herman need really have looked no further than that.

That completes my case that Burks, in particular, promulgated the von Neumann myth. If I am correct in this, then it seems fair to add also that Burks' particular adoption of the myth would have been decisive for its subsequent development (given his authoritative position as the editor of the relevant von Neumann manuscripts), and that is why I have discussed his case in such detail. Difficulties which then flowed from this can be summarised relatively briefly.

Since von Neumann's original development, his results are been rederived in a variety of different frameworks. I include here, for example, Thatcher's (1970) redesign of a UGM within von Neumann's original space; Codd's (1968) work on a "simpler" 8-state space; Berlekamp, Conway & Guy's (1982) outline work on a 2-state space; and Arbib's (1969b, Chapter 10) formulation which shifted back somewhat toward the kinematic kind of model. This is probably by no means an

[26] Granted, Herman does work with Codd's definition of UCM, which I consider deeply misleading, as already explained; but that does not affect the application of his argument to Burks' claim.

exhaustive list. I hold that, whatever other merits this kind of work might have had, it has not offered any advance in terms of von Neumann's original problem. In particular, there has been no recognition here of the substantially modified problem situation which resulted from solving P_v. And I blame this, in large measure, on the von Neumann myth: if von Neumann's *original* problem is not understood, or mistaken, the *new* problem situation will also be missed.

That the myth is still alive and well is apparent from, for example, Langton (1984). Langton, as Herman before him, senses that there must be something wrong with the myth. Langton's version of the myth seems a little more garbled—he cites the embedding of a UCM as a criterion for non-triviality in self-reproduction, but he may mean this in Codd's sense (which refers, in effect, to a universal computational power, and is thus ultimately related to Burks' version). In any case, Langton stipulates that this criterion is not satisfactory for various reasons. With this I agree whole-heartedly. But in contrast to Herman (assuming they are talking about essentially the same thing) Langton feels that the criterion is too *strong* rather than too weak. He therefore goes on to propose, as a replacement criterion, that we require only that self-reproduction involve separate processes of "copying" and "decoding" a description. In this way he manages to preserve the superficial form of von Neumann's analysis, while cutting out its heart; for Langton describes an automaton which still has a vaguely von Neumann-like mechanism of self-reproduction; but in which the description language has been so impoverished that there are absolutely *no* A-mutations which would yield another, different, but still self-reproducing, automaton. On my interpretation this must be seen as a cruel (though of course unintentional) parody of von Neumann's work, which could not possibly have been proposed if von Neumann's true problem (rather than the myth) had been properly understood. It is all the more ironic when viewed in the light of a subsequent paper (Langton 1986), when this intrinsically deficient (by Von Neumann's criterion) self-reproducing automaton is again described, but this time followed by an extended discussion of the possibility of a Darwinian evolution process among self-reproducing automata—a discussion in which Langton fails entirely to recognise the deep problems which this raises, at least some of which von Neumann had long before not only recognised but solved!

To conclude this discussion: I assert that there *is* a von Neumann myth, which seriously mistakes the nature of the problem which von Neumann confronted (and solved); that it is pernicious and persistent; and that it has seriously hampered, if not completely preempted, further progress in the direction of realising artificial Darwinism. I emphasise again that my criticism here is not at all directed at the people who have subscribed to the myth: it is purely of the objective myth itself. I believe that it has caused considerable damage, and that is why I have felt justified in expending so much effort on its identification, elaboration, and refutation.

I hope that I am correct in my analysis; and, if so, that the myth can now finally be dispelled.

4.3 A New Problem Situation

4.3.1 P_a: The Problem of *Autonomy*

Von Neumann's formulation and solution of *some* of the fundamental problems underlying the (Darwinian) growth of complexity in formal (or 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 (U)GM 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)?

The first point to make in relation to this is that, as far as I am aware, it has never been empirically tested. Indeed, not even the operation of a single A-reproducer on the von Neumann design has been so tested. According to Kemeny (1955, p. 66) Von Neumann's basic A-reproducer would occupy about 200,000 cells (a size dominated by what I have called the A-descriptor, which stretches out for a linear distance of about 150,000 cells). Thus, to implement[27] a large enough example of this A-system to support not just a single A-reproducer, but a sufficient *population* of such A-reproducers that they may interact and form competing S-lineages—and thus to potentially allow for Darwinian evolution—would be a very daunting task. Matters would not be dramatically better for the alternative cellular A-systems of, say, Codd (1968) or Berlekamp *et al.* (1982); although the individual cells are simpler in the latter systems, the size of configuration required to realise a von Neumann style A-reproducer would be (more or less) correspondingly larger.

The second point to make is that 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,[28] 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

[27]There is, perhaps, room for argument about the meaning of "implement" in this context—specifically, whether a "simulation" on, say, a conventional, serial, computer would qualify. However, I consider that to be a sterile essentialist argument, and will not take it up. In this particular case, the reader is invited to adopt whichever meaning she prefers; it will not materially affect the conclusions.

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

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, 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. However, I do *suspect* this may be the case, simply due to the fragility of the underlying cell states—they can typically

be disrupted by almost any perturbation. Again, von Neumann suggested as much, commenting that this may be, in part at least, "the price one has to pay for the simplicity obtained by our elimination of kinematics" (von Neumann 1966b, p. 130). I may say that, in this respect, the cells of von Neumann's original cellular A-system, though more complicated than those of other cellular A-systems subsequently proposed, were certainly more robust—much closer, in this respect, to von Neumann's informal kinematic A-system. However this, on its own, would surely not suffice to make the *basic*, genetic, A-reproducer(s) even in von Neumann's cellular A-system, as described by von Neumann, Burks and Thatcher, genuinely viable; and this situation could only be worse for those A-systems where the underlying cell states are individually more fragile.

Having said that, I do not wish to lay any great stress on this issue of the fragility or otherwise of the primitive *cells* (or, more generally, A-parts) in an A-system. I fully accept the general conclusions from, for example, Langton's (1986) extensive review of this question in the particular context of cellular A-systems. To paraphrase very roughly, it will only be if there is some kind of compromise ("balance" is Langton's word) between fragility and, we may say, rigidity, in the properties of the A-parts that the existence of A-machines having a wide variety of A-complexity will be possible at all. My point however is that there may be an almost literal danger here of missing the wood for the trees. While we certainly need some kind of suitable "trees" (A-parts of appropriate potentialities), this by no means automatically solves the problem of building a "wood" (viable, robust, A-reproducers).

Thus we may say that designing "good" A-parts seems like a step in the right direction—but it is a step of unknown size, and it *might* be exceedingly small compared to the journey ahead. 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 what I am calling the problem of *autonomy*; and I venture to suggest that it is much the harder problem.

I may also mention here the `VENUS` system 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; in particular, we cannot directly introduce

the idea of Genetic Pluralism. 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.

The problem P_a may thus be stated 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). A-reproducers satisfying these condi-

tions could, I suggest, be reasonably termed *A-organisms*.[29]

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 biological 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 many other things I have discussed, and will yet discuss. 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 &

[29] I mean that this is "reasonable" only in the sense that it seems not to do *too* much further violence to the English language; but, of course, I should not be read as making any metaphysical claims for having finally, definitively, isolated the one true *essence* of life here. A word is a word is a word.

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 the von Neumann myth has, to some extent inhibited work on P_a; 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. With this background, I shall then finally formulate a suggestion for a particular kind of *indirect* attack, which will serve to conclude the chapter.

4.3.2 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.[30]

"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. None of this is intended as any criticism of Holland himself, for (as far as I can see) he has never claimed that the Genetic Algorithm *did* solve P_a. Indeed, although I state my argument in the specific context of the Genetic Algorithm,

[30] I shall introduce quite a different suggestion of Holland's, the so-called α-Universes, in the concluding section of this chapter.

the fact that it is really directed at the underlying problem situation rather than at this particular attempted solution means that it should be taken to apply *mutatis mutandis* to a variety of other work also.

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.

4.3.2.1 Holland's Problem (P_h)

I have already reviewed the underlying philosophical commitments of Holland and his colleagues (Holland *et al.* 1986) in the previous chapter (section 3.8.3). I concluded there that the processes which Holland *et al.* describe as *inductive* are, precisely, processes of *unjustified variation* in the sense of UVSR; 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 I am here calling *Holland's problem* or P_h, and I recognise it as a genuine and difficult problem.

The important point for my purposes is this: the growth of knowledge requires two things—unjustified variation *and* selective retention (reflective of "verisimilitude"). P_h concentrates almost exclusively on the former, whereas P_a concentrates almost equally exclusively on the latter. My problem (encapsulated in P_a) is not concerned at all with the rival "merits" of different heuristics or generators or sources of variation (though it requires that some such sources of variation must exist); rather it is concerned almost exclusively with selection mechanisms—indeed, with one particular selection mechanism, that of Darwinian *natural selection*.[31] I am not arguing here for some preeminence of either

[31]There is, as always, no claim that, for example, Darwinian, natural, selection, is *guaranteed* to select for "verisimilitude"; merely that it sometimes *might*, and is, moreover, the best, if not the only, example we know.

problem—the growth of knowledge relies on at least partial solutions to *both*; I merely hope to have established that they *are* distinct.

4.3.2.2 P_v Again...

I contend that P_v can 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 de-coupling; 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 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; explicit comments on this appear in (von Neumann 1966a, p. 84) and (von Neumann 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 the point at issue in the Genetic Algorithm is emphasised by other elements of the problem situation which underlay Holland's work. As noted in the previous chapter, 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. We shall see that this is, at least roughly, the idea of the Genetic Algorithm.

However: the crucial point, for my purposes, 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 *selection for verisimilitude*. Indeed, it does not even identify selection as a problem. Conversely, selection *is* the substantive new issue being raised in P_a. Thus, whereas P_h takes selection as relatively unproblematic, and concentrates on variation, P_a takes variation as relatively unproblematic and concentrates on selection (specifically, natural selection).

Still: this argument does not yet quite make P_h and P_a *disjoint*. In particular, it does not *necessarily* mean that the Genetic Algorithm is, as I claim, 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. Therefore, I shall now briefly outline the Genetic Algorithm, comment on how it can, perhaps, be regarded as a partial solution to P_h, but then show how it is hardly relevant to P_a.

4.3.2.3 What is the Genetic Algorithm?

Suppose that there exists a population of entities, which Holland calls *structures*, but which, for my purposes, will be equated with A-reproducers. Suppose that, associated with each such A-reproducer there is an A-descriptor, in the sense of a data storage subsystem whose contents remain essentially static for the lifetime of any single A-reproducer, and which establish (describe) the complete structure and organisation of that A-reproducer. Associated with each A-reproducer there must also a measure of its "degree of adaptation", which Holland normally calls *fitness*; I shall take this to be equivalent to A-knowledge in my terms.

The Genetic Algorithm may then be described as follows:

1. Arrange (somehow) that the total population size is limited to some maximum value.

2. Arrange (somehow) that the A-reproducers do, indeed reproduce; but that, furthermore, the relative reproductive success of each A-reproducer is proportional to its A-knowledge. That is to say that if we think, roughly, in terms of discrete generations, the expected relative number of surviving offspring for any A-reproducer will be proportional to its relative A-knowledge.

3. Arrange (somehow) that, in the process of reproduction, the A-descriptors are subject to certain specified kinds of transformations, or "genetic operators". These would include something essentially equivalent to what has previously been termed A-mutation, but would also include something akin to recombination in biological organisms. Holland refers to the latter as a crossover operator; I shall call it *A-crossover*. It denotes the construction of an offspring A-descriptor by splicing together segments taken from two distinct, parental, A-descriptors. The use of some form of A-crossover is the most distinctive characteristic of the Genetic Algorithm.

4.3.2.4 What good is the Genetic Algorithm?

The Genetic Algorithm preserves (implicitly), from the prior solution of P_v, the notion of A-descriptors as passive subsystems, which can therefore be used, via the definition of the description language, to configure A-machines in general, and

A-reproducers in particular, into a more or less arbitrary genetic space, having the property that A-reproducers are close together in this space.

Indeed, applications of the Genetic Algorithm are commonly arranged so that *only* A-reproducers inhabit the genetic space—i.e. an arbitrary transformation of a point in the space is guaranteed to yield another A-reproducer. This corresponds, in the von Neumann model, to disallowing A-mutations (or any other kind of genetic transformations) affecting those parts of the A-descriptors coding for the core machinery (g_0): essentially, attention is restricted to that part of the A-descriptor coding for the "ancillary" machinery ($x \in X$). Von Neumann's work mandates this kind of assumption in the sense that von Neumann showed (by concrete example) that a descriptor language could be implemented which allowed A-descriptors to be factored or decomposed in this way. However, it is worth noting that to adopt this view is tantamount to adopting Genetic Absolutism; it is therefore a somewhat restrictive decision, as discussed in section 4.2.6 above.

In any case, the key novelty which the Genetic Algorithm introduces is that transformations of the A-descriptors are no longer limited to the kind of local A-mutations envisaged in von Neumann's schema, but are now expanded to include A-crossover. A-crossover allows relatively "large" transformations to be tried out in genetic space. The significant difference between A-crossover and simply increasing the A-mutation rate (per A-part in the A-descriptor—which would ultimately allow similarly large transformations) is that the transformations to be tried are severely constrained. Roughly speaking, only points which are a cross between existing points will be sampled via A-crossover. The conjecture is that, in many cases of practical interest, this kind of transformation will be "better" than any comparable kind of A-mutation, in terms of the A-complexity of the transformed A-reproducers.

Of course, this is not the whole story. The Genetic Algorithm introduces what I have elsewhere (McMullin 1992a) called *bimodal* procreation—the idea that a single offspring has multiple parents. This, in turn, allows intersecting S-lineages, and means that a number (possibly a large number) of S-lineage selection processes can go on concurrently within a single population. Holland has placed considerable emphasis on this point, referring to it as *intrinsic parallelism* (Hol-

land 1975) and/or *implicit parallelism* (Holland 1986). In explaining this Holland introduces the concept of a *schema*, being a set of A-descriptors which are "identical" in certain specified respects; it is essentially identical to Dawkins' (1976) notion of a "selfish gene", and corresponds, in my terms, to a tag identifying a particular S-lineage. Holland's point is then that any single A-reproducer will be an element of many schemata, and thus its reproductive success (or otherwise) can simultaneously contribute to many different S-lineage selection processes.

I have previously argued, at length, that, in the presence of epistasis, the operation of this kind of concurrent selection may become problematic (McMullin 1992c, esp. section 7.2.1). This is particularly so if selection involves Sewall Wright's process of *shifting balance* (e.g. Wright 1982). It seems to me suggestive that at least one application of a form of the Genetic Algorithm (Mühlenbein *et al.* 1988) actually involved deliberate modifications of the population structure which were very reminiscent of the conditions required for a shifting balance process to operate. Mühlenbein has recently made this connection with the Shifting Balance process more explicit (Mühlenbein 1992).

However, be that as it may, it is not central to my concerns here. Let us accept that intrinsic parallelism may be a significant and useful effect. This will be most obvious in the case that there is little or no epistasis; and in that case (at least) the operation of intrinsic parallelism can be viewed as involving the independent, concurrent, selection of relatively short segments of A-descriptors (which are largely undisturbed by A-crossover) which, as they come to dominate the population, are automatically joined together (by the operation of A-crossover). This is the so-called "building block hypothesis" concerning the operation of the Genetic Algorithm (Goldberg 1989, 41–45); situations (such as mentioned in the previous paragraph) in which this hypothesis may not hold are then generally referred to as *GA-deceptive* (Goldberg 1989, pp. 46–52). The point, for my purposes, is that, although the idea of intrinsic parallelism is overtly associated with selection, its force is concerned with its advantages (if any) for the *generation* of new variation.

That is, even allowing for the operation of intrinsic parallelism, the Genetic Algorithm is strictly concerned with the problem P_h (the problem of *generating* variation) rather than with P_a (the problem of *selecting* variation). P_a is not con-

cerned at all with the selection "dynamics" as such; it is concerned with selection *criteria*; and these are not addressed at all by the Genetic Algorithm (in itself). Somewhat the same point has been made previously by, for example, Mühlenbein (1989). The point is manifest in my particular formulation of the Genetic Algorithm in the previous section, where it is simply *stipulated* that reproductive success (and thus, eventually, selection) *is* conditioned by A-knowledge—without any comment on how this can be achieved in practice.

None of this rules out the possibility that a particular *application* of the Genetic Algorithm *might* address P_a. Since every such application must involve *some* selection criteria, these *may* be the kind of criteria sought by P_a. As it happens, I am not aware of any such applications: selection is typically performed relative to a "fitness" function, which may be explicit or implicit, static or dynamic, but which ultimately reflects criteria established by the researcher rather than criteria emerging spontaneously within the A-system itself (i.e. they do not incorporate *natural* selection). In other words, whatever growth of knowledge occurs in these systems is parasitic upon, and constrained by, the prior knowledge of the researcher.

But even if some application *did* address P_a in this way, my point is that it would not be doing so *by virtue* of incorporating a Genetic Algorithm; its relevance to P_a would, rather, be an essentially independent attribute. I conclude that the Genetic Algorithm, interesting though it may be in its own domain, has nothing to offer in the solution of P_a.

4.3.3 Constraining the Interactions

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: there can be no selection at all in these systems, never mind selection for verisimilitude. To put it another way, 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). I note that this work is relatively recent, and its publication postdates the rest of the analysis presented in this chapter. My discussion of `Tierra` is therefore limited to a preliminary review, sufficient only to assess its effect on my central conclusions.

`Tierra` can roughly viewed as a development of the `VENUS` system discussed in section 4.3.1 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 Pluralism 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 yet be said to 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.

4.3.4 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 1973] 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 simi-

larity 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 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. I shall not discuss it further at this point, but I will eventually return to it in the next chapter (section 5.5.8).

4.4 Conclusion

The major purpose of this chapter has been to reconsider and reinterpret von Neumann's work on Automata Theory. The result is a claim that the problem which von Neumann was primarily concerned with was, precisely, that of Artificial Darwinism—the growth of knowledge in artificial systems by Darwinian mechanisms. Conversely, and contrary to the received wisdom, I claim that von Neumann was *not* interested in the "problem" of self-reproduction as such, but only in the connection of this problem with Artificial Darwinism. Furthermore, von Neumann was able to provide an important part of a solution to the latter problem. The key element of this was to show how, in almost any "reasonable" axiomatization of automata theory (i.e. which is strong enough to support fairly general notions of computation and construction), there can exist large and diverse sets of A-reproducers whose elements are connected under some plausible idea of A-mutation. This is achieved by introducing the idea of a self-reproduction architecture based on A-descriptors, which largely decouples mutational connectivity from the specific structures of the A-machines.

Arising from this result, I identified a new problem, denoted P_a. This is, roughly, the problem of how A-reproducers, of von Neumann's general architecture, can be sufficiently robust to actually carry out their reproductive function in a more or less *hostile* environment. Alternatively, we may say that P_a is concerned with exhibiting a set of A-reproducers, spanning a wide (preferably infinite) range of A-complexity/A-knowledge, which can *practically* support the operation of natural selection. This is an informal, and still rather poorly defined problem (though its formulation can be significantly improved through the

introduction of the concept of *autopoiesis*). But I argue that, even in this crude form, P_a is of central importance; and that little substantive progress has yet been made toward its solution.

In conclusion, I want to suggest a new strategy, or research programme, 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!).

My 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). Holland provided some initial theoretical analysis of his proposal, but he then left the idea aside. In the next chapter therefore, I shall revisit this proposal of Holland's, and report on a detailed empirical investigation.