Artificial Knowledge

An Evolutionary Approach

Finbarr Vincent McMullin

A thesis submitted for the degree of Ph.D.

OLLSCOIL na hÉIREANN
THE NATIONAL UNIVERSITY OF IRELAND

University College Dublin
Department of Computer Science

October 1992

Head of Department: Professor F. Anderson
Supervisor: Dr. John Kelly

Dedication

For Colette: partner, wife, lover, mother of our children—but, more important than any of these, a true and enduring friend.

Contents

Abstract

I present a new analysis of the *problem situation* in Artificial Intelligence (AI), grounded in a Popperian epistemology.

I first review arguments purporting to establish that no purely "computational" system can realise *genuine* mentality. I conclude that the question is still open; but that the more pressing question is whether such a system can even exhibit intelligent *behaviour*. Attention is thus directed at the computational embodiment of *knowledge*, and its *growth*. I suggest that much of the work in this area incorporates a flawed, naïve empiricist, epistemology. I adopt Popper's view that the growth of knowledge is possible *only* through a process of *unjustified variation and selective retention*. In particular, the inate knowledge of biological organisms has arisen by such a process, in the form of Darwinian evolution.

I review previous work on the realisation of Darwinian processes in computational systems. In particular, I present a critical reinterpretation of von Neumann's pioneering work in this area. I conclude that no system to date has exhibited substantive growth of artificial knowledge via a process of Darwinian evolution. More importantly, I argue that this problem is deeper than is generally recognised, requiring the effective integration of *autopoiesis* with *evolvability*. To achieve this it may ultimately be necessary to realise something analogous to the *genesis of life*. I review one proposal for such a phenomenon: Holland's so-called α-Universes. I present an *implementation* of a specific α-Universe and review the (largely negative) results of empirical tests carried out on it.

I conclude with the claim that the problem of realising the spontaneous genesis of "artificial life" is of great difficulty, but that its solution may yet prove to be an essential prerequisite for the realisation of anything deserving to be called "artificial intelligence".

Acknowledgements

Whatever there may be of value here, it was made possible only through the most generous help from many different individuals and organisations. I am grateful to all who have helped me over the years.

I must thank my parents firstly: they gave me, and my brothers Paul, Paddy, Michael and Brian, the great gifts of freedom and opportunity, while somehow managing to ensure that we did not abuse them. I should do so well for my own children!

In my education and professional life I have been helped by very many people; but I must mention, in particular, Jim Lacy who let me become an Engineer in my own way; Michael Healy and Pat O'Donoghue who actually paid me to do it; Mark Dargan, Brendan Curtin and Brendan McMahon who helped me have fun doing it; Stephen Brennan who told me about HAHL while it was a fun place to be; Sean O'Riordan who let me sign up, and still let me go on doing things my own way; all the gang who built the original *autonomous systems* (complete with the famous *berserk mode*)—Michael Brogan, Jon Bruty, John Clancy, Nada Couzi, Arthur Cremin, Frank Devitt, Joe Fitzpatrick, James Kelly, Paul McHugh, Joe McManus, Ray Moran, Ciaran Mulloy, Eileen Noonan, Robert O'Dea, Niall O'Donnell, Bridget Walsh—to name only a very few (if we ever have another reunion, it will have to be in the Holiday Inn, Irvine); and Charles McCorkell, of DCU, for giving me yet another chance to go my own way.

In my five years in Dublin City University, I have benefited greatly from discussions with very many colleagues both in DCU and elsewhere. I am indebted to all the staff of the School of Electronic Engineering in DCU for continuing encouragement, even as their eyes glazed over; Noel Murphy and Paul Whelan have been especially tolerant, far above and beyond the call of duty. Paul McKevitt made useful comments on earlier drafts of the Thesis, while a Research Fellow in DCU. A special thanks is due to my uncle, Ernan McMullin, for sending me the Daedalus, Winter 1988, special issue on AI: it was more help than he can have imagined; thanks also for resisting the temptation to tell me how far wrong I was probably going. Gabriel McDermott, then in UCD, helped greatly when I was trying to find a starting point. John Holland of the University of Michigan, Chris Langton of Los Alamos National Laboratory, Steen Rasmussen of the Technical University of Denmark, Tom Ray of the University of Delaware, and Peter Cariani of the State University of Binghampton were all most generous in responding to my correspondence. Francisco Varela of CREA, Paul Bourgine of CEMAGREF, and John Stewart of CNRS, provided very helpful advice also.

The following are people I've never met or communicated with, but, in the most literal possible sense, this Thesis could never have appeared on paper without them. So, thanks to Don Knuth for TEX; Leslie Lamport for LATEX; Oren Patashnik for BIBTEX; Eberhard Mattes for emTEX; Tomas Rokicki for `dvips`; David Rhead for the author-date style file; Stephen Page for double-spacing etc.; and Niel Kempson for `MS-DOS` BIBTEX.

DCU has been generous and enlightened in its material support, in particular paying the NUI examination fee, and all fees associated with my enrollment in University College Dublin. And speaking of UCD, John Kelly has been the very model of a Ph.D. supervisor, giving me great enormous lengths of rope, and still, incredibly, managing to stop me hanging myself.

Thanks also to all my family-in-law, who long since adopted me as one of their own.

Emma Jane McMullin was born just as I was setting out on the road that would eventually lead to this tome; David John joined in a little later. They have enriched my life immeasurably. They never understood why Daddy spent so much time locked up in his "study", but they took it with extraordinarily good humour, and never ever interrupted.

Well, almost never.

I have dedicated this work with all my love to my wife, Colette McNamee: absolutely above and beyond everything and everyone else, it simply could not have happened without her.

Prologue

All this happened in a small town in Ireland, some time ago.

There was a man who went to live in the woods, thereby greatly surprising his neighbours. He was a well set-up man with a good job and a fine house, and so was well thought of, but he sold the fine house and gave away the money for it, and he gave up the good job. Now this was a time when a good job was hard to come by and a man didn't give one up unless he had a better one to go to, or he was an eejit. The man who went to live in the woods had never been thought of as an eejit, but now his neighbours wondered. They asked him why he was doing this thing.

"I don't know," said the man who went to live in the woods. This was thought to be an unsatisfactory answer, so they concluded that he was an eejit after all, and left him to his folly...

Brendan McNamee
The Man Who Lived In Sorcy Wood

Chapter 1

Setting Out

1.1 Introduction

The journey which this Thesis involves is a somewhat intricate one. While each separate chapter is reasonably self-contained, and might be read in isolation, the essential thrust of the work relies on the interconnections between them. The purpose of this introductory chapter is therefore to preview the major landmarks which will appear along the way, and especially how they are related to each other. Equipped with this outline the reader will then hopefully be in a position to examine the details without losing sight of the overall view.

1.2 On Criticism

> The point I want to make here is that Popper's work itself contains a feature, unavoidable when properly understood, which has got between him and potential readers—who, being only potential, are not yet in a position to understand it. He believes, in a sense which will be made fully clear later, that only through criticism can knowledge advance. This leads him to put forward most of his important ideas in the course of criticizing other peoples' ...
>
> Magee (1973, p. 14)

I am very far indeed from supposing that anything I present here would bear favourable comparison with the achievements of Karl Popper. But Magee's comment is relevant in at least this one respect: this Thesis is quite deliberately and self-consciously a work of *criticism*. I believe that I have some new things to say, but that the only way to say them is to place them securely in the context of the

problems they attempt to solve. These are not new, so to present the problems means to revisit the work of their originators; and to offer new solutions means to criticise previous solutions, and to show where, in my view, they are deficient and can be improved. I emphasise this at the outset, for otherwise the reader may quickly find herself wondering when I am going to stop merely "reviewing" the work of earlier writers, and start with my own substantive contribution; Gentle Reader, do not look for this boundary for it is nowhere to be found. My "substantive contribution" *is* precisely this critical review, and cannot be conveniently distinguished from it.

1.3 Popper's Problem

> I, however, believe that there is at least one philosophical problem in which all thinking men are interested. It is the problem of cosmology: ***the problem of understanding the world—including ourselves, and our knowledge, as part of the world.*** All science is cosmology, I believe, and for me the interest of philosophy, no less than of science, lies solely in the contributions which it has made to it. For me, at any rate, both philosophy and science would lose all attraction if they were to give up that pursuit.
>
> Popper (1980, p. 15, original emphasis)

I shall call this problem of cosmology *Popper's Problem.* I do not, of course, propose to solve it. Indeed, I have very little to say *directly* about it. Nonetheless, I think it worthwhile to make explicit, this once, the fact that it is the original motivating problem which will be lying behind the various more specific problems with which I shall be visibly concerned in this work.

My approach to this problem of Popper's is inevitably conditioned by my training as an Engineer. The first instinct of the Engineer is to take things apart, and the second is to put them together again—only differently. That is, as an Engineer I try to understand by re-creating. I don't expect to re-create the world, and, in truth, I don't really expect to understand it. But I might succeed in understanding some bits of it; the trick is to select those bits which are interesting, *and* for which there is some realistic chance of success in understanding, which is to say in re-creating.

1.4 Making a Mind

The bits of the world with which I shall start are *minds.*

It is an obvious, if rather foolhardy, starting point. Popper's Problem is thus exchanged for something which is not noticeably any easier: the problem of *Artificial Mentality*—building or re-creating minds. It is the subject of Chapter 2.

I consider only one relatively narrow aspect of this more specific problem: whether we can establish valid *a priori* grounds for rejecting one particular approach—namely the attempt to realise an artificial mind *simply by executing an appropriate program on a digital computer.* The latter, which I shall call the hypothesis of *computationalism*, may be taken to be the premise underlying the research programme of *Artificial Intelligence* (AI) at least in its so-called "strong" form (Searle 1980).

I have two points to make about this.

The first is that the idea that computationalism is true is an affront to human dignity. However this does not make computationalism false. More importantly, even if I "believed" that it is true, this would be at best a *fallible* belief—I would not use it, in itself, to undermine a humanist ethics.

My second point is that I do not accept the arguments put forward either by Searle (1980) or Popper & Eccles (1977) for rejecting computationalism on strictly *a priori* grounds. This, of course, does not make computationalism true.

I finally wash my hands of this problem, by saying that I am a metaphysical dualist (I really and truly believe that computationalism is false); but that I am simultaneously a methodological computational monist (I am going to pretend that computationalism is true, because that seems, currently, like the most promising avenue for making any progress).

1.5 Knowledge and Its Growth

With the really difficult problems thus held in abeyance, I address myself to something which is at least superficially much more tractable. Let us not aim to realise a computational *mind*; instead, we will settle for computational *knowledge.* This amounts to asking for a computer to exhibit *behaviours* which we characteristi-

cally associate with mentality, while we withhold judgement as to whether this could ever be the "real" thing. The problem of realising artificial, particularly computational, knowledge is therefore tackled in Chapter 3; this is the problem of Artificial Intelligence in the "weak" form (Searle 1980).

I take up a number of current issues in AI, which, though they are quite distinct, are not independent: there is a single objective motivating my entire discussion, which is the attempt to strip away the considerable clutter and verbiage that has accumulated in the vicinity of the modern AI research programme, and to thus lay bare what I consider to be its bedrock: the problem of the *growth* of artificial knowledge. I suspect that many workers in the field are not even clearly aware of the existence or true nature of this problem; and are certainly not aware of its depth and difficulty.

My first sally here is concerned with the so-called *Turing Test* (Turing 1950). This is an operational, or behavioural, "test" for intelligence, based deliberately and exclusively on linguistic performance; it has provided an important focus for AI research. I have two comments to make about it. First, the Test has recently been criticised by French (1990) for being too stringent; I attempt to clarify the nature of the Test, and to show, in this way, that French's criticism is unfounded, and that the Test can still serve as a valid goal in AI research. However, secondly, and more importantly, I suggest that it is, at best, a very *long range* goal; no computers have come close to passing this test, and there is little immediate prospect that any will. It seems to me wildly premature to actively pursue this specific goal in the current state of the art. In my view, effective linguistic performance relies on very substantial pre-linguistic knowledge; I suggest that, for the time being at least, attempts to achieve linguistic behaviours are a distraction from the real problems confronting AI.

I turn next to the vexed question of "cognitive architecture": roughly, what "kind" of computer is "best" for realising AI? This is a question which implicitly underlies much of the tension between the two major contemporary groups within AI: those advocating the "symbol processing" or "Good Old Fashioned AI" (GOFAI) approach, and the "connectionists". I will not preview that discussion in detail here: suffice it to say that I consider this debate to be futile. There is, of course, *no* "best" kind of computer for realising AI; and discussion in those terms

is, again, a distraction from the real problems.

At this point, I digress to attempt to clarify what it is I mean by "artificial knowledge". Briefly, I equate knowledge with the generation of predictions about the world, which are at least "approximately" true, and the exploitation of these predictions to effectively mediate an agent's interaction with the world. Knowledge thus consists in anticipatory models or expectations, and is relative to the world in which the agent is embedded. There is, perhaps, nothing shockingly new in this view, but it contrasts with some of the ideas typically entertained within AI, and it is worth spelling out for that reason.

With this more precise concept of "knowledge" in hand, I consider the problem of embodying such knowledge in a computer system. I argue that doing so with a brute force, so-called *knowledge engineering*, approach is unsatisfactory for two reasons. The first is pragmatic: the experience has been that this is an extremely difficult thing to so. In itself this is not decisive—perhaps we simply have not yet tried hard enough. The second reason for rejecting knowledge engineering is, on the other hand, fundamental and compelling: we should rightly consider any system which relies on this form of spoon feeding—which is incapable of "learning" for itself—as a peculiarly impoverished and unsatisfactory kind of "intelligence". Thus finally do we expose what I have already called the bedrock problem of AI: the *growth* of artificial knowledge.

How one progresses beyond this point depends critically on a *philosophical* issue: the problem of *induction*. Strangely, this need for a definite epistemological foundation rarely seems to be made explicit in AI; that is to say, many workers in AI seem not to recognise that there *is* any "problem" of induction (e.g. Lenat 1983). Be that as it may. I adopt the Popperian view, which is simply that there is no such thing as a "logic" of induction; but that, notwithstanding this, knowledge can and does grow by a kind of generalised Darwinian process of *unjustified variation and selective retention* (UVSR).

I review this theory of evolutionary epistemology, and point specifically at the distinction between "unjustified" variation (a strictly logical notion) and "unbiased" variation (which is a quite different notion concerned with *verisimilitude*). I argue that several apparent criticisms of the UVSR approach to knowledge growth rest on a confusion between these two notions, and are therefore unfounded.

At this stage the problem at hand has resolved itself into the following form: can we build a computational system which can support an open-ended growth of knowledge, based on the principles of Popperian evolutionary epistemology?

I may note that Popper himself has been less than sanguine about the prospects for such a development:

> We learn by mistakes; and this means that when we arrive at inconsistencies we turn back, and reframe our assumptions. In applying this method we go so far as to re-examine assumptions even of a logical nature, if necessary. (This happened in the case of the logical paradoxes.) It is hardly conceivable that a machine could do the same. If its creators, incautiously, equip it with inconsistencies, then it will derive, in time, every statement that it can form (and its negation). We may perhaps equip it with a gadget which will warn it, in case it derives '0=1', and make it abandon some of its assumptions. ***But we shall hardly be able to construct a machine which can criticize and readjust its own methods of derivation, or its own methods of criticism.***
>
> Popper (1988, p. 109, emphasis added)

This comment originally dates from about 1957, and could perhaps be criticised for being over-simplistic in the light of developments in automated logic since that time. Nonetheless, I think the crucial point, contained in the final sentence which I have italicised above, still stands: it raises the problem of making the system *self-referential* in a very deep and fundamental way. This remains a very difficult and intractable problem, and lies behind much of what is to be discussed in this Thesis. In my own earliest analysis, I described such systems as being *reflexive*, and summarised the difficulty like this:

> We have now exchanged an abstract philosophical problem for a (mere?) engineering one: how to actually design and build such reflexive systems. More carefully: it is easy to design a system which is reflexive—the problem is that it will tend to immediately self-destruct. This phenomenon is familiar to all who have had programs "accidentally" treat their own instructions as data, and overwrite themselves—a "crash" is the inevitable result. Thus we need to identify what properties or constraints a reflexive system should have so that it will spontaneously evolve toward greater internal organisation, and correspondingly sophisticated external behaviour. In short, a system which, even if not initially intelligent, can *become* intelligent.
>
> McMullin (1990, p. 214)

In any case, at this stage in the discussion the problem of the growth of knowledge has been recognised as continuous with, and in a certain deep sense,

identical with, the problem of the growth of organismic complexity through Darwinian evolution. Given that Darwinian evolution is the best concrete example of evolutionary epistemology in action, I now reformulate the problem in the following way, finally taking it altogether out of the conventional domain of "artificial intelligence": can we abstract the processes of Darwinian evolution from their biological source, and embody them in a computational system?

1.6 On Darwinism

It seems to me that any serious attempt to realise an artificial, computational, Darwinism, should best be preceded by a serious analysis of the nature of Darwinian theory within its original, biological, domain. However, I have presented such an analysis in detail elsewhere (McMullin 1992a; 1992b; 1992c), and I will not repeat that material here. Instead, I proceed directly to the question of embodying Darwinian evolution in a computational system, and this is the subject for Chapter 4.

While there has been a recent resurgence of interest in this issue (particularly under the rubric of *Artificial Life*—e.g. Langton 1989a; Langton *et al.* 1992; Varela & Bourgine 1992), the seminal work was carried out by John von Neumann in the period 1948–1953 (von Neumann 1951; Burks 1966d). I present a detailed re-evaluation and critique of von Neumann's work.

My first, and perhaps most important, point is that von Neumann was indeed concerned with the realisation of an artificial Darwinism in a computational medium. This requires emphasis, and detailed argument, because there has emerged what I shall call a von Neumann *myth* in this area, which suggests something quite different. The myth holds that von Neumann was working on some problem of automaton "self-reproduction" *per se*; and because this would admit of trivial "solution", the myth further holds that von Neumann introduced, as a criterion of automaton "complexity" (and thus of "non-trivial" self-reproduction), a requirement that a universal computer (or, perhaps a "universal constructor") should be embedded within it.

Like all myths, there is a core of truth in this; but the myth is now very garbled, and the truth is extremely hard to uncover.

Briefly, I argue that von Neumann was interested in the question of automaton self-reproduction only insofar as that is an element of the problem of realising artificial Darwinism; and that, insofar as he proposed a criterion for "non-trivial" self-reproduction, this was simply that it should be such that it *can* potentially support the growth of automaton complexity by Darwinian processes. Von Neumann's genius was firstly to recognise that this is problematic at all (he pointed out that it seems paradoxical that any automaton could construct another which is more complex than itself) and secondly that this very particular problem can be overcome by using a kind of *programmable* constructing automaton (what I shall call a *Genetic Machine*).

These points, once they are distilled, are, it seems to me, fairly clear and uncontroversial. However, the detailed arguments are rather involved and will take up the bulk of Chapter 4.

The balance of that chapter is concerned with going beyond von Neumann's work: he had solved one important aspect of the problem of realising artificial Darwinism, but this by no means represents a complete solution. Von Neumann showed how one could design an automaton such that it could, *in principle*, construct other automata more complex than itself (and so on). In practice, however, von Neumann's design can work only if his automaton is protected from virtually all manner of interference or perturbation of its operation—conditions which effectively rule out any possibility of Darwinian natural selection taking place. In this way, the outstanding problem, not addressed by von Neumann, is identified as the problem of *autonomy*—how can an automaton establish, maintain, and protect its own unity and integrity in the face of environmental perturbations. It is, in its way, a *prior* problem, and perhaps a deeper and more difficult one.

I examine a range of work which may be said to have been inspired, directly or otherwise, by von Neumann's investigations.

I suggest that a significant portion of this work is contaminated by the von Neumann myth—for if one has adopted that mistaken view of von Neumann's original problem, it becomes almost impossible to see, much less to solve, the outstanding problem of autonomy. A rather different criticism may be levelled at another indirect offspring of von Neumann's work—the so-called *Genetic Algorithm* (e.g. Holland 1975; Goldberg 1989). I shall argue that the Genetic

Algorithm is concerned exclusively with the rival merits of different processes of "unjustified variation" which might be overlaid on a basic von Neumann style artificial Darwinism. This is, no doubt, an interesting issue in its own right, but it is, at best, tangential to the problem of autonomy.

The problem of autonomy *has* been directly confronted by some researchers, most notably Humberto Maturana and Francisco Varela (Maturana & Varela 1980; Varela 1979). They have formulated an explicit and technical notion of autonomy, which they call *autopoiesis*. Roughly, a system is autopoietic if it is self-regulating or homeostatic in respect of its own identifying organisation. Furthermore, Varela *et al.* (1974) have demonstrated artificial autopoietic systems within a computational framework which is at least loosely inspired by von Neumann's work. However, while these workers have demonstrated artificial autopoiesis, the systems they exhibit are no longer self-reproducing—not, at least, in the strong sense of a von Neumann style genetic self-reproduction.

The problem of achieving a growth of artificial, computational, knowledge (or, what amounts to the same thing at this point, a growth in automaton complexity, in von Neumann's sense) now seems to amount to this: can we embed, in a suitable computational framework, automata which are autonomous in the sense of autopoiesis, and which also satisfy the von Neumann conditions for a Darwinian open-ended growth in complexity?

As far as I am aware, this problem has not been solved; indeed, it is unclear whether it has been previously recognised as an important problem in its own right. While it may well be that this problem will eventually succumb to a direct attack, I choose instead to consider the possibility of an indirect attack. This suspends the attempt to directly build or engineer systems which would satisfy the desiderata set out above, and asks instead whether such phenomena might *spontaneously* arise under suitable conditions? In biological terms, we redirect our attention away from the *evolution* of life, and take up, instead, the question of its *genesis*.

1.7 The Genesis of Artificial Life?

I take up the question of *Artificial Genesis* in Chapter 5, but I do so in a rather narrow and specific way.

Recall that we are interested in the spontaneous emergence of entities which are both autonomous (autopoietic) and satisfy the von Neumann conditions for an evolutionary (Darwinian) growth of complexity—what I shall loosely call *evolvability*. The first of these seems not too difficult—it has been specifically exhibited by Varela *et al.* (1974). In fact, I believe that the phenomenon has been encountered in other systems also, though not generally recognised as such; this, for example, is the sense in which I would interpret Ray's otherwise fantastic remark that "It would appear that it is rather easy to create [artificial] life" (Ray 1992, p. 393). In any case, combining the spontaneous emergence of autopoiesis with von Neumann's conditions for evolvability is another, and altogether more difficult, problem.

As it happens, there has been at least one specific attempt to formulate systems which would specifically support the spontaneous emergence of self-reproducing entities using a von Neumann style genetic mechanism: these are the α-Universes introduced by John Holland (1976). It should be emphasised that Holland's proposal was made in a rather different context from that in which I attempt to apply his work. In particular, even if the α-Universes did everything which Holland thought they might, this would not represent a solution to the problem which *I* have formulated. Firstly, although the self-reproducing entities envisaged by Holland use a kind of genetic mechanism, they still fall far short of satisfying the von Neumann conditions for evolvability (they do not span a significant range of "complexity"). Secondly, Holland did not address the issue of whether these entities would be autonomous at all—not, at least, in the technical, autopoietic, sense. However, having said that, the implication of Holland's analysis was that these entities would be at least "viable" in face of a range of "perturbations", and thus it seems that the α-Universes could provide a useful stepping stone toward solving the problems at hand.

Holland provided a description of one specific α-Universe, which I denote α_0, and his detailed theoretical analyses were based specifically on this. His results

were concerned with estimating the expected spontaneous emergence time for primitive genetically self-reproducing entities; specifically he proposed that this could occur in a somewhat incremental fashion, and that this could make the difference between a feasible and a totally infeasible emergence time.

Holland did *not* carry out any empirical testing of his results, though he noted that it should be possible to do so. The bulk of my Chapter 5 is therefore concerned with presenting just such a programme of empirical testing. This involves firstly re-defining α_0 in considerably more detail, and with greater formality, than Holland's original presentation; the latter left many details open, which was satisfactory for Holland's purely theoretical purposes, but such details must be specified in any practical implementation. I then review the results of a series of tests of Holland's predictions.

The outcome of this is, in effect, a report on failure. It turns out that Holland's analysis was flawed, insofar as it neglected several significant effects which he had not anticipated. α_0 cannot, in fact, support the predicted spontaneous emergence of genetic self-reproduction, not even in Holland's relatively impoverished sense of that.

Failure however, is not necessarily a *negative* outcome. While α_0 does not behave as expected, the precise modes of failure are interesting, and may provide a useful basis for further work. In particular, having investigated α_0 in detail, it becomes clear that, at best, it could only ever have realised autopoietic entities in essentially the same, rather limited, sense as had already been implicitly exhibited by, say, Rasmussen *et al.* (1990) or Ray (1992). That is, the entities which would be autopoietic would *not* be the putatively genetically self-reproducing entities. Once this is recognised, it suggests some possible avenues for further exploration, which would attempt to *combine* relevant aspects from these several different systems, and also from the somewhat different systems of Varela *et al.* (1974) and Zelany & Pierre (1976).

However: such further investigations would finally take us beyond the scope of what can be addressed in this one Thesis, and must therefore be left simply as aspirations for the future.

1.8 Conclusion

To conclude this introductory chapter I shall summarise once more.

I have been inspired by Popper's great cosmological problem of understanding the world and our place within it; but I know that that problem is too demanding, and so I immediately simplify by focusing attention on understanding ourselves, and what kinds of things we might be—which is to say the problem of *mentality.* Here, and throughout, I seek an engineer's solution—by re-creating we might understand. I examine some of the arguments against the very possibility of a computational re-creation of mentality, but conclude that they are not compelling. Now I simplify again, leaving aside mentality proper (that ineffable notion of self-conscious experience) and ask whether we can re-create intelligent behaviours, which is to say *artificial knowledge.* I attempt to strip away various ancillary issues which have come to obscure this problem, and argue that the substantive issue is then the *growth* of artificial knowledge. Further progress demands certain philosophical commitments, and I commit myself to a Popperian evolutionary epistemology. I thus simplify again, and now ask whether we can re-create a form of *artificial Darwinism.* I show how von Neumann solved one important aspect of this problem; and how this leaves exposed another, perhaps more basic and more difficult aspect, namely the re-creation of artificial *autonomy.* My final simplification is to ask not for artificial Darwinism, but for artificial *genesis.* I describe one detailed attempt to achieve this; and conclude by examining, and trying to learn from, its failure.

That is the journey ahead. We can do no more planning; we must simply set out.

Chapter 2

Artificial Mentality

2.1 Introduction

> Many psychologists and brain scientists are embarrassed by the philosophical questions, and wish no one would ask them, but of course their students persist in asking them, because in the end these are the questions that motivate the enterprise.
>
> Dennett (1978b, Introduction, p. xiii)

> In coming to grips with the idea of a natural system, we must necessarily touch on some basic philosophical questions ... This is unavoidable in any case, and must be confronted squarely at the outset of a work like the present one, because one's tacit presuppositions in these areas determine the character of one's science.
>
> Rosen (1985a, p. 45)

This chapter is concerned with the philosophical milieu in which the rest of the Thesis will be unfolded. More particularly, it is concerned with the question of whether the research programme which goes under the title of *Artificial Intelligence*, or *AI*, is capable (even in principle) of solving any of the substantive problems posed by the existence of *minds*.

This is no idle concern. As we shall see, a variety of critics, most notably Searle and Popper, have suggested that the answer to the question is a more or less simple *No*—that AI *cannot* illuminate the problems of mentality. If they were correct in this assessment it would represent a limitation, at the very least, on the applicability of subsequent discussions in the rest of the Thesis. This is so because, although I eschew many of the conventional tools and techniques associated with AI research, the work I describe still falls within the essentially

computational paradigm which identifies AI as a field. It is as well to confront this issue at the outset.

My objective then, is to confound at least some of the critics of AI.

Having said that, let me immediately emphasise that my conclusion will be the weakest possible in the circumstances: I claim merely that the case against AI, or "computationalism" in the broadest sense, is *not (yet) proven.* It is quite enough for my purposes that the question still be open. Specifically, I do not propose to argue that AI demonstrably *can* solve any particular problem(s) of mentality. Or, if you wish, I accept that the case *for* AI (as an approach to mentality at least—I ignore any questions concerning technological *utility*) is, equally, not (yet) proven.

I suspect that this agnostic position is implicitly shared by most workers in AI; however, as Rosen points out in the quotation above, it is best to be explicit about such preconceptions.

2.2 Three Hypotheses

I shall state three related hypotheses, which will then serve as targets for criticism.

H_p **(Physicalism):** All mental states and events can, in principle, be completely reduced, without residue, to physical states and events.

H_c **(Computationalism):** All mental states and events can, in principle, be completely reduced, without residue, to computational states and events, of some universal computer.

H_t **(Turing Test Computationalism):** The Turing Test (Turing 1950) can be passed by certain systems whose *putative* mental states and events can, in principle, be completely reduced, without residue, to computational states and events, of some universal computer. H_t is, essentially, a behaviouristic version of H_c.[1]

[1] I shall review the Turing Test in detail in the next chapter. For present purposes, the following formulation is adequate: a system passes the Turing Test if, based on purely linguistic interrogation (e.g. via teletype), but spanning arbitrary topics, a competent judge mistakes it for a person.

H_c implies both H_p and H_t. Thus the following scenarios are logically conceivable:

- A refutation of H_c would be neutral with respect to both H_p and H_t.
- A refutation of H_p would be neutral with respect to H_t, but would constitute a *de facto* refutation of H_c.
- Similarly, a refutation of H_t would be neutral with respect to H_p, but would constitute a *de facto* refutation of H_c.

H_c is the hypothesis of direct interest in this chapter; I have introduced H_p and H_t solely because any (alleged) refutations of these would also refute H_c.

There are, of course, many other relevant hypotheses closely related to those I have introduced here, but with varying flavours and technicalities. However, in general, I deliberately overlook such finer distinctions in what follows, because they seem to be unnecessary refinements for the relatively modest purposes I have in mind.

2.3 A Personal Bias

> ...Yet machines are clearly not ends in themselves, however complicated they may be. They may be valuable because of their usefulness, or because of their rarity; and a certain specimen may be valuable because of its historical uniqueness. But machines become valueless if they do not have a rarity value: if there are too many of a kind we are prepared to pay to have them removed. On the other hand, we value human lives in spite of the problem of overpopulation, the gravest of all social problems of our time. We respect even the life of a murderer.
>
> Popper & Eccles (1977, Chapter P1, p. 4)

Before proceeding to consider criticisms of H_p and H_c, I should like to declare an element of personal bias: I side with those who hold that physicalism, whether in the plain form of H_p, or the more specific form of H_c, is utterly and irredeemably repugnant to human values. I shall therefore digress briefly to document just why I continue to regard physicalism with such distaste.

2.3.1 *Why* Physicalism is (Still) Repugnant

Physicalism is repugnant because it denies the freedom and responsibility of man.

This is hardly a novel or original view, though it may have become less fashionable to speak of it (in the context of AI, at least). Indeed, some will, no doubt, consider me naïve to persist in it. However, I believe that this view, though hackneyed, is essentially correct. It has been effectively argued as such by, for example, Popper (Popper 1965; 1988; Popper & Eccles 1977).

Briefly, the physicalist hypothesis may be viewed as equivalent to the claim that the physical world is causally *closed* (which is, of course, not at all the same thing as claiming that the physical world is *deterministic*). This being so, mental states and events (i.e. minds, as such) can, in principle, be *dispensed* with in any description or analysis of physical states and events.

Minds may, of course, still be convenient devices for summarising certain (physical) phenomena. That is, minds may usefully be deployed in describing certain "law-like" physical behaviours. Indeed, it seems to me that it could only be by virtue of some such fact that minds, like thunderstorms or galaxies, would be real entities in good standing at all (regardless of the truth or otherwise of physicalism). But, even at best, the physicalist position is that any description of states and events, which incorporates *mental* states and events, will be exactly equivalent to some alternative (albeit vastly more complicated) description in terms purely of *physical* states and events. Indeed we might expect mentalistic descriptions to be mere approximations to the purely physicalist descriptions (though this is not crucial to the argument).[2]

In particular, consider any episode of the (apparent) exercise of "freedom"—that is, some kind of rational, or at least considered, decision making. If physicalism is true then, in principle, the initial set of mental states (and any other relevant factors) can be reduced to physical states; the trajectory of the system

[2]Both Smolensky (1991) (with his plea for the "Proper Treatment of Connectionism" or PTC), and Hofstadter (1979; 1983) (with his concept of "tangled hierarchies"), have given interesting discussions of such an *approximate* relationship between mentality and physics. A detailed review would take me too far afield here, but I briefly consider Hofstadter's views again in section 2.3.2 below.

can be evaluated by reference only to these physical states;[3] and the physical result or outcome (which, in the general, stochastically indeterministic, case will not be unique, but will rather be represented by a probability function or distribution) can then be encoded back into the resulting mental states, which will represent the decision (or a probability function or distribution over potential decisions).

This is, of course, simply a restatement of Laplace's thought experiment which envisaged a "demon" who could know the instantaneous dynamic state of the entire universe, and could therefore predict the entire behaviour of the universe for all future time. The only additional feature I have introduced is to allow for stochastic or probabilistic rather than strictly deterministic dynamics—in deference to the stochastic form of quantum mechanical physical theories. This does not, in any way, affect the force of the argument with regard to the exercise (or not) of human "freedom".

Note carefully that the argument does not rely at all on the *practicality* of a Laplacian demon. In particular, although Popper (1988) has provided a variety of arguments against what he terms " 'scientific' determinism", this latter doctrine is much stronger than the mere causal closure of the physical world claimed by H_p. " 'Scientific' determinism" seems to require that a Laplacian demon be physically realisable, at least in principle (I take this to be implicit in Popper's "principle of accountability" and his requirement for prediction from *within* the physical world). Whereas, in the discussion of "freedom" above the point is not whether the future can, in fact, be predicted (statistically or otherwise), but whether it can be *altered* (statistically or otherwise). In this respect, H_p is much closer to what Popper calls "metaphysical determinism", a doctrine implied by, but much weaker than, " 'scientific' determinism".

In any case, the essence of H_p is that no alternative analysis of the genesis of a human "decision", using mentalistic terms or otherwise, could say any *more* about the relationship of that "decision" to the prior state of the universe; indeed, we expect that a mentalistic analysis would yield, at best, only a poor, and

[3]There may be some difficulty with establishing what constitutes the "system" here; but, if needs must, we allow this to include the entire physical universe (regardless of whether this is bounded or unbounded). Recall that this is only an "in principle" discussion.

incomplete, approximation to the physicalist result. In short, mental states and events would have to be considered as, in some sense, *epiphenomenal.*[4] It is true that, under H_p, the outcome of a decision may not be deterministic (i.e. be a *unique* function of the prior state), but it cannot be reasonably said to be the "free choice" of the person; the possible outcomes, and their relative probabilities were already determined, and were not changed one iota by the particular thoughts that the person (appeared) to think.

The loss of freedom implied by H_p carries with it, of course, the loss of responsibility or moral obligation: since the person's thoughts (desires, intentions etc.) can be dispensed with in evaluating her actions, we could hardly hold her responsible for those actions.

Taken to its logical conclusion of course, this signifies that my very discussion of this topic is also epiphenomenal, and, in that sense, ridiculous (though perhaps not quite absurd). This result is, in essence, what Popper has termed "the nightmare of the physical determinist" (Popper 1965, p. 217) because it takes its clearest form under the hypothesis of a deterministic physical universe, in which a unique trajectory property holds. However, the point which Popper was at pains to expose is that *the nightmare is not in the least relieved by a stochastically indeterministic physics.* As long as a complete reduction of mental states and events to physical states and events is possible, in the sense that the resulting description is causally closed (whether the causation is deterministic or stochastic) then the nightmare recurs, as I have tried to make clear above. In Popper's words, "indeterminism is not enough" (Popper 1965; 1973).

2.3.2 Some Contrary Views

There exist, of course, a variety of contrary views on the repugnant consequences of physicalism.

Firstly, a common supposition is that stochastic indeterminacy (typically, though not necessarily, involving an appeal to quantum mechanics) can make physicalism and human freedom compatible. Indeed, this was the view of Arthur

[4] "Epiphenomenalism" comes in more than one flavour. The kind I have in mind here is that of Hofstadter (1983)—which seems to be subtly different from that of Popper (Popper & Eccles 1977, Chapter P3, Section 20).

Holly Compton, as noted by Popper in his Compton memorial lecture (Popper 1965); however, Popper firmly rejected this view, essentially for the reasons discussed in the previous section. The point is that physical indeterminacy of this sort simply does not change the nature of the argument, nor, therefore, its conclusion. I shall not consider this position further.

A.F. Huxley has argued that the *impracticability* of actually carrying out a physicalist reduction robs it of its sting:

> ...I used to be upset at the idea of possibly not having free will, but it now seems to me that even if we do not have free will, the events which govern our movements are so unpredictable that there is no need to be worried about it.
>
> Huxley (1983, p. 15)

Penrose has recently offered a more sophisticated variation on this argument, in the context of a discussion of free will. He argues that a complete physicalist analysis of any particular mental event or events may be not merely impractical, but actually *impossible*, in the technical sense of being *uncomputable* (Penrose 1990). In this way, the physical world could, in fact, be causally closed, but in such a way that this closedness could not be exploited from within.

In a sense, however, the relatively sophisticated appeal which Penrose makes to the notion of computability is unnecessary. The following argument, formulated by, for example, Popper (1974c, Section XXV), seems to me to establish the same point more directly and decisively. Reality, in its entirety, is causally interconnected (by definition). Thus, any *practical* attempt to make a complete analysis of any aspect of reality (from within the real world) would require a complete model of the real world to be embedded within itself; this would, of course, include a model of the model, and so on. This is clearly impossible (incompletable).

So let it be stipulated that a *complete* reduction of mentality to physics will always be impractical; the point remains that the repugnance of physicalism rests entirely on its *in principle* nature, and not any particular claim to be able to carry it out; the latter would be a factor in any attempt to *corroborate* physicalism, but that is not the issue just here. Indeed, Penrose himself seems to finally acknowledge that the impracticality of a physicalist reduction cannot, in itself, restore human freedom to the universe (Penrose 1990, pp. 558–559).

Another possible position is to accept the consequences of physicalism, but to put a brave face on the situation—claim that it may not be intrinsically repugnant after all.

Sperry put this view succinctly when he said "There may be worse fates than causal determinism" (Sperry 1965, p. 87). It should be stressed that Sperry does not mean *strict* determinism here—he specifically accepts that a stochastic *indeterminism* would add nothing more than a degree of "unpredictable caprice" to our actions. Rather, he is referring to the general physicalist position that mental events are ultimately reducible to physical events, which is to say H_p.

However, Sperry's position is still a good deal more complex than the slogan might suggest. Hofstadter (1985, Chapter 25) has provided an extended allegory expanding on this paper of Sperry's. Ultimately, in fact, both Sperry and Hofstadter seem to be ambivalent about the implications of physicalism for free will. That is, as far as I understand them, they adopt physicalism, accept that this is incompatible with "free will" as it is conventionally understood, and yet they also seem to qualify their physicalism, as if to draw back again from this abyss.

Thus, Sperry claims, in effect, that we can have our physicalist cake and eat it:

> ...you will note that the earlier basic distinction or dichotomy between mentalism and materialism is resolved in this interpretation, and the former polar differences with respect to human values ...become mainly errors of reductionism. This may be easily recognised as the old "nothing but" fallacy; that is, the tendency, in the present case, to reduce mind to nothing but brain mechanism, or thought to nothing but a flow of nerve impulses.
>
> ...Our quarrel is not with the objective approach but with the long accepted demand for exclusion of mental forces, psychic properties, and conscious qualities—what the physicist might class as "higher-order effects" or "co-operative effects"—from the objective scientific explanation.
>
> Sperry (1965)

Like Sperry, Hofstadter emphasises the existence of "emergent" behaviours, in the sense of levels of description having their own distinctive characteristics, even though these are still "compatible with" (but does this mean "reducible to"?) a purely physical level of description. In the case of conscious experience and free will, Hofstadter particularly emphasises the Gödelian implications of self referential symbols at different levels of description (Hofstadter 1979, Chapter XX).

It seems to me that both Sperry and Hofstadter are here confusing two quite different issues: the *utility* of mentalistic (or other "higher-order") descriptions, versus their *necessity.*

It is certainly the case that there exist descriptions of states and events in the world, incorporating mentalistic terms, which are approximately, if not exactly, true—indeed, it was stipulated in the previous section that this is actually the *defining condition* (at least in a causally closed physical world) for such "higher-order" entities to be recognised at all. These mentalistic descriptions, being "higher-order", are more concise and tractable than the corresponding purely physical descriptions. This is enough to make them useful additions to, or even replacements for, purely physical descriptions, for practical purposes of analysis and prediction.

But none of this implies that "higher-order" (mentalistic or otherwise) descriptions are *necessary.* Indeed, the point of saying that the physical world is causally closed is precisely to say that non-physical entities are, even if only in principle, superfluous to a complete account of physical states and events.

To be fair to Sperry and Hofstadter, neither explicitly claims that their approach does anything to restore the dignity of man in a soulless universe. At the end of the day, they are more concerned with reinterpreting our *attribution* of free will, than in restoring or rehabilitating the real thing. This is a perfectly sensible procedure upon the adoption of a physicalist position, but I do not see that it can make physicalism in the least degree more *palatable.*

There is a final possible position to be considered, though it really brings us full circle. This is to claim, *despite* the arguments marshalled in the previous section, that physicalism somehow *is* compatible with the exercise of free will, and the attribution of responsibility. This is a position which Dennett forthrightly promised to defend:

> ...can psychology support a vision of ourselves as moral agents, free to choose what we will do and responsible for our actions? Many have thought that materialism or mechanism or determinism ... threaten this vision, but ... *I consider the most persuasive of the arguments to this effect and reveal their flaws.*
>
> Dennett (1978b, Introduction, p. xxii, emphasis added)

However, virtually in the same breath, we already find a partial retreat from this bold and intriguing promise:

> By uncovering the missteps in the most compelling arguments for this thesis *I claim not to refute it, but at least to strip it of its influence.*
>
> Dennett (1978b, Introduction, p. xxii, emphasis added)

I shall ultimately find myself more or less in agreement with Dennett in this *second* formulation; that is, while I continue implacably to assert the repugnance of physicalism, I agree that this need not "influence" our scientific investigation of it. However, as I shall discuss in the next section, my grounds even for this circumscribed position are somewhat different from, and more general than, Dennett's.

But, before proceeding to that, I should like to comment briefly on the detailed arguments which Dennett actually presented (Dennett 1973).[5]

Dennett primarily argues for the validity of adopting what he calls the *intentional stance* toward certain systems, specifically including people. This is a necessary step in his argument since the intentional stance is, he says, "a precondition of *any* moral stance, and hence if it is jeopardized by any triumph of mechanism, the notion of moral responsibility is jeopardized in turn" (Dennett 1973, pp. 242–243).

This is all true, but is not, in my view, *germane*. There is no doubt that the intentional stance can usefully be adopted in many situations, and that this possibility is a requirement for intentional systems, like minds, to be recognised as such at all. But it is not clear that *anyone* is arguing to the contrary (i.e. to the effect that the ultimate *truth* of purely physical description would, in some sense, imply the *falsity* of mental, or intentional, descriptions). The point is not that mentalistic or, more generally, intentional, descriptions are false, or even useless, but rather that they may be causally redundant. The physical world might, as it were, go along just the same way without them.

The notion that physicalism might somehow rule out the adoption of the intentional stance, for *utilitarian* purposes, is a distraction—a mere straw man. That

[5] Dennett has since provided a much more extensive analysis of "free will" and related problems (Dennett 1984). Chapter 5 of that work addresses the issues of most concern for my purposes, but I have been unable to identify anything which would deflect the criticisms which I present of Dennett's earlier essay (Dennett 1973). A properly comprehensive review of (Dennett 1984) would take me too far afield; I shall therefore not discuss it further.

is, as long as the intentional stance is merely that—a "stance" we might choose to take up with respect to certain physical systems, for utilitarian purposes—it seems that it cannot be relevant to the issue under discussion here.

However, Dennett does offer a few further twists that might affect this conclusion. He considers the point that to abandon the intentional stance toward *oneself* would be fundamentally incoherent; that there is therefore an element of intentionality in the world which *is* more than an optional stance toward an essentially physical system—for the very taking up of a stance is, in itself, an intentional action.

This is an intricate and intriguing argument. But Dennett himself immediately admits that it is really an attempt to refute physicalism, rather than a means of reconciling physicalism with free will. And, as a refutation of physicalism, it fails. Briefly, it is another deterministic nightmare: if physicalism is true, we cannot properly be said to "choose" to take up any stance at all.

Popper has discussed this kind of argument critically, and provides the following concise version of what can, and cannot, be validly drawn from it:

> ... the epiphenomenalist argument leads to the recognition of its own irrelevance. This does not refute epiphenomenalism. It merely means that if epiphenomenalism is true, we cannot take seriously as a reason or argument whatever is said in its support.
>
> Popper & Eccles (1977, Chapter P3, p. 75)

Indeed, I may say that this analysis provides the rationale for the entire orientation of the current chapter: I consider the arguments *against* physicalism, but not those *in favour*; for, by definition, the most compelling arguments in favour of physicalism must also be the most self defeating.

But to return to Dennett, he next considers the point that:

> ...no information system can carry a complete true representation of itself ... And so I cannot even in principle have all the data from which to predict (from any stance) my own future.
>
> Dennett (1973, p. 254)

But this is simply back to the question of the *practicality* rather than the *truth* of physicalism; indeed, Dennett explicitly acknowledges Popper's formulation of this point, as I have already described it in the discussion of Huxley and Penrose above; and it still does not impinge on the issue of free will.

It seems then that Dennett does not achieve his original aim of showing how free will and physicalism might genuinely co-exist in a single cosmology. His concluding remarks are, in fact, addressed to a different theme:

> Wholesale abandonment of the intentional is in any case a less pressing concern than partial erosion of the intentional domain, an eventuality against which there can be no conceptual guarantees at all.
>
> Dennett (1973, p. 255)

The issue is no longer the relationship between free will and physicalism; but rather the potential for abuse of whatever physicalist understanding of mentality (if any) may, in practice, be achieved. This is now a discussion of the *uses* of science, which is to say a *moral* discussion. As such, it is not, itself, any longer a part of the scientific discourse. This is the point at which I can finally agree with Dennett, and I elaborate this general position in the next section.

2.3.3 But: does it *really* matter?

> ...Thus I regard the doctrine that men are machines not only as mistaken, but as prone to undermine a humanist ethics. However, this very reason makes it all the more necessary to stress that the great defenders of that doctrine—the great materialist philosophers—were, nevertheless, almost all upholders of humanist ethics. From Democritus and Lucretius to Herbert Feigl and Anthony Quinton, materialist philosophers have usually been humanists and fighters for freedom and enlightenment; and, sad to say, their opponents have sometimes been the opposite.
>
> Popper & Eccles (1977, Chapter P1, p. 5)

Does it matter that the physicalist hypothesis has dehumanising implications? Well, the fear expressed above by Popper, that it might be used as an excuse for dehumanising *actions*, is not entirely without foundation. Perhaps this explains, in part, why a proponent of physicalism might be loath to accept that this position does indeed imply the abandonment of human freedom and responsibility. If so, this would be quite understandable, perhaps even admirable in its way; but I suggest that it would also be quite mistaken.

We can and should face up to the consequences of our theories, *even* when they are odious. We can do this because, in fact, there is nothing to fear from even the most odious consequences of any theory—*provided we remember that our theories are just that, fallible inventions of the human mind.* I cannot accept that

any such fallible theory, no matter how well corroborated, could ever provide us with *moral* principles or, worse, *justifications.* "Scientific morality" is, I suggest, a contradiction in terms. As Popper has said, even the greatest defenders of physicalism have actually been upholders of the humanist ethic—and that, in my view, is precisely as it should be.

In short, I assert that it is only good science to admit the implications of our theories, repugnant or not; *but that it is then only good philosophy to admit that our scientific theories, in themselves, are devoid of moral authority.* Science absolves no sins.

In this present context, this means that the implications of physicalism, repugnant as they may be, can still be viewed with a certain degree of equanimity or detachment. We might almost say that this makes the attribution of repugnance aesthetic rather than scientific; as such, it need not, and should not, deflect the scientific investigation.

Having said that, I should emphasise that I do not suggest that the path of science is free from moral decisions, or from moral culpability—that scientific "progress" might be justified as an end in itself. Quite to the contrary, I consider that scientific activities are no different from any other human activities in this respect; they share the moral imperative for us to consider (as well as possible), and accept responsibility for, the likely outcomes of our activities. It is precisely in discharge of this moral obligation that I have stipulated my abhorrence of physicalism, *per se*, have positively argued that this abhorrence is justified, but have then gone on to argue that this, in itself, does not have the force of a general, moral, restraint on the scientific investigation of physicalist theories of mentality.

This position should be distinguished from, say, a specific advocation of scientific investigation into theories of "brain-washing, subliminal advertising, hypnotism and even psychotherapy ... and the more direct physical tampering with drugs and surgical intervention" (Dennett 1973, p. 255). Such activities could, no doubt, fall within a physicalist research programme; but, as Dennett implicitly draws out, they would require specific moral validation well beyond anything which has been discussed here.

2.4 Refuting Computationalism?

Having made clear the unhappy implications of H_p (and thus H_c), but having also affirmed that this should not, in itself, deflect us from the further scientific study of these hypotheses, I now return to the substantive question of this chapter: *Has H_c already, in fact, been refuted?*

One avenue for the attempted refutation of H_c is the claim, originally propounded by Lucas (1961), that Gödel's results on the existence of undecidable propositions in consistent formal systems establish that mentality is necessarily irreducible to formal processes. However, this has already received extensive exploration, and many detailed criticisms (see, for example, Hofstadter 1979; Dennett 1970; Hofstadter & Dennett 1981, p. 470, p. 475 give further references). I shall therefore make only one brief comment here.

The Lucas argument relies on the claim that, faced with any machine which putatively exhibits mentality, one can always formulate a proposition which the machine cannot prove but which any *person* (Lucas himself, for example) can see to be "true". As Dodd (1991) has pointed out, albeit in a slightly different context, any such perception of "truth" is actually dependent on an assumption of *consistency* for the relevant formal system; but, precisely because of Gödel's results, such consistency *cannot*, in general, be proven. Thus, it seems to me that the argument by Lucas fails from the very start: while the machine cannot prove its Gödel sentence, *neither can Lucas*; the most that Lucas can do is to *conjecture* that it is true—and I can see no bar to the machine also doing that much. In any case, I shall not pursue the Lucas argument further.

I now turn to two other, quite distinct, arguments for the refutation of H_c. These are Searle's so called *Chinese Room* thought experiment, and the rather more general "dualist interactionist" argument for the causal openness of the physical world (which is to say, for the falsity of H_p, and thus, implicitly, of H_c also) presented by Popper & Eccles. It seems to me that these are substantial and challenging arguments, and I shall devote the following sections to considering them in some detail.

2.4.1 Searle's Chinese Room

> Searle's [Searle 1980] 'Chinese Room' argument against 'Strong AI' has had considerable influence on the cognitive science community ... it has challenged the computational view of mind and inspired in many respondents the conviction that they have come up with decisive, knock-down counterarguments ... Yet the challenge does not seem to want to go away ... Indeed, some have gone so far as to define the field of cognitive science as the ongoing mission of demonstrating Searle's argument to be wrong.
>
> Harnad (1989)

John Searle's original presentation of his Chinese Room argument was already accompanied by extensive peer commentary (Searle 1980). In the twelve years that have since passed, there has been a continuing stream of publication on the issue. A survey is provided by, for example, Harnad in the paper quoted above. Slightly more recently, *Scientific American* has hosted another instalment in the debate, with a restatement of his position by Searle, and an attempted rebuttal by P.M. Churchland and P. Smith Churchland (Searle 1990; Churchland & Churchland 1990). It is clearly a matter of some continuing interest and significance for AI, and I should therefore like to comment on it.

In what follows, I shall take "Strong AI", as Searle terms it, as being equivalent to my H_c, and "Weak AI" as equivalent to my H_t.

Searle's contention is that H_c is false, and that this is demonstrable through a series of thought experiments. I shall describe only the simplest of these, and even that only very briefly.

Let there be a computer which (when suitably programmed) appears to instantiate the mentality of a Chinese speaking person (in something like the sense of the Turing Test). A person, ensconced in the so-called *Chinese Room*, could, given appropriate, purely formal, instructions, simulate the behaviour of this computer exactly. This Chinese Room would also, therefore, putatively instantiate the mentality of the Chinese speaking person. The "real" person carrying out the simulation is stipulated not to be a Chinese-speaker. If we now enquire of this person whether she understands any Chinese, she will say no. Therefore (?) there is no genuine Chinese mentality being realised by the Chinese Room, and therefore mentality cannot be reduced, without residue, to computational states and events. H_c has been refuted.

It is important to note that Searle *accepts* H_p, or, at least, something essentially equivalent to it:

> **Can a machine have conscious thoughts in exactly the same sense that you or I have? If by "machine" one means a physical system capable of performing certain functions (and what else can one mean?), then humans are machines of a special biological kind, and humans can think, and so, of course machines can think. And, for all we know, it might be possible to produce a thinking machine out of different materials altogether—say, out of silicon chips or vacuum tubes. Maybe it will turn out to be impossible, but we certainly do not know that yet.**
>
> Searle (1990, p. 20)

So, Searle's claim is that some sort of physicalist (H_p) theory is (or at least, may be) true—but that H_c is not that theory.

Searle is neutral with respect to H_t: indeed, the Chinese Room argument only works given the assumption that H_t may, in fact, be true (if H_t somehow actually proves to be false, then that automatically refutes H_c anyway, and the fact that the Chinese Room argument could no longer even be properly formulated would not matter—it becomes redundant with respect to the real problem, i.e. the truth or otherwise of H_c).

Now most, if not all, commentators on this issue can be divided into two groups:

- Those who hold that H_c is false, whether they agree with all of Searle's reasoning or not. Thus I include here, for example, Eccles (1980), who agrees with Searle's refutation of H_c, but disagrees strongly with Searle's uncritical acceptance of H_p (Eccles describes himself, following Popper, as a "dualist interactionist"—see Popper & Eccles 1977; I shall consider their views in more detail in section 2.4.2 below).

- Those who hold that H_c is true. Their basic position is that, since H_c is true, Searle *must* be wrong. They then go on, *in the light of this,* to try to identify precisely why Searle is, in fact, wrong. I consider that, if *any* of these particular commentators are right, it is those who advocate the so-called "systems reply". Briefly, this grants that the person in the Chinese Room *per se* does not have any Chinese understanding or mentality, but holds that the Room *as a systemic whole* (including the person inside)

understands, or at least, might understand, Chinese—i.e. have "genuine" Chinese mentality. However, I shall not pursue the arguments for and against that position here.

My purpose in making this classification is to identify, by omission, a third possible position: that which holds that Searle's reasoning is wrong, and that, therefore, the status of H_c is simply *unaffected* by his argument: it remains a tentative hypothesis. This is the position I propose to adopt.

It is important to realise that this is a perfectly valid procedure, and is, if correct, preferable to a position of claiming that H_c is actually true. It is preferable in the basic sense that attempting to argue for the truth of the converse of a proposition is, in general, an *unnecessarily strong* way of attacking a supposed proof of the original proposition. But the procedure is doubly preferable in this particular case where any attempt to prove the truth of H_c inevitably undermines itself anyway (it is another variant of the "deterministic nightmare" of section 2.3.1). I suspect that this may be at the root of Harnad's observation that, "Many refutations [of Searle's argument] have been attempted, but none seem convincing" Harnad (1989, p. 5).

So, to reiterate, my claim is that Searle's reasoning is defective, and his conclusion (that H_c is false) is therefore *unwarranted*; but I do *not* suggest that H_c is, in fact true. My only claim is that its status is still open.

Briefly, the argument is this:

> H_c does not make the prediction which Searle ascribes to it (that the person in the Chinese room should, upon enquiry, report that she understands Chinese); in fact, H_c is entirely neutral as to the outcome of the experiment. H_c cannot, therefore, be *refuted* by Searle's experiment—*no matter what its outcome!*

As far as I am aware, this argument is due, in essence, to Drew McDermott, who introduced it in personal communication with Harnad; I have not identified any published version of precisely this idea. In my view, this argument is not only concise and elegant, but also devastating. On the other hand, as Harnad stated in my opening quotation above, many have previously thought they had identified "decisive" arguments on this issue, but the debate rumbles on nonetheless (indeed,

Harnad himself rejected this view of McDermott's, but I have been unable to understand his reasons).[6]

In any case, I now turn back to Searle's own arguments. Searle has, I think, been somewhat puzzled by the reception his ideas have had—at least in the AI community. He believes that his Chinese Room Argument is decisive against H_c, and yet there are many people who are unwilling to accept this. So he seeks an explanation of this. He finds a candidate explanation in the notion that some people may (mistakenly) think that H_t necessarily implies H_c. Therefore, anyone who accepts Turing's original argument for H_t (basically, a universal computer can realise any effective procedure—can "simulate" anything whose behaviour is sufficiently well specified—and there is no manifest *a priori* reason for supposing that human linguistic performance cannot be so specified) would interpret this as an argument for H_c also; and might therefore be convinced that Searle must be wrong in his refutation of H_c, even if they cannot identify exactly *why* he is wrong.

Now even Searle himself is willing to accept the *possibility* that H_t may be true. So he perceives that part, at least, of his task should be to show how it can be that H_t could be true, and yet H_c could be false.

He does this by citing other phenomena (e.g. rainstorms) which can be perfectly well *simulated* by computers, but which plainly cannot be so *realised* (a simulated rainstorm cannot make you wet!). By analogy, he argues, there is no reason to suppose that the mere simulation of a mind (H_t) would actually cause a "real" mind to be called into existence (H_c)—(Searle 1980, p. 423).

My comment is simply to say that all this is certainly true, insofar as it goes, but it is not germane; at least, it is not germane to *my* disagreement with Searle.

Thus, I *do* say that, in a certain special sense, H_t *might* imply H_c; but this is not my reason for rejecting the Chinese Room Experiment, and it is not at all affected by spurious meteorological analogies (ironically, Searle himself warns against the dangers of wanton analogising—Searle 1990, p. 24). In fact, the situation is exactly opposite to that apparently envisaged by Searle.

[6]Excerpts from this correspondence between Harnad and McDermott were distributed by Harnad through his electronic discussion group on the so-called *symbol grounding problem*; my discussion is based on a message dated `Sun, 13 May 90 23:11:40 EDT`.

I *start* with a rejection of the Chinese Room argument (following McDermott, as explained above). I therefore also, implicitly, reject Searle's alleged distinction between mere mind-like behaviour (H_t) and real minds (H_c). I then conjecture that, in the absence of some alternative criterion for distinguishing H_c from H_t (i.e. independently of the Chinese Room Experiment) the two are (*pro tem*) identical (i.e. the Turing Test is a *bona fide* test for mentality); and in this very special, degenerate, sense, it can actually be technically correct, although not very illuminating, to say that H_t implies H_c (rainstorms notwithstanding).

Or to put it another way, Searle's analogy only begins to make sense if we already accept that minds are entities like rainstorms, whose realisation demands certain specific, physical, causal powers, and are *not* entities like computers (or, if you prefer, computations) which can be realised by more or less arbitrary physical systems; but if we already accepted *that,* we would have already accepted the falsity of H_c, and the analogy would be unnecessary. It seems that, whichever way you look at it, Searle's discussion of simulation versus realisation does not add anything to the original argument.

Of course, on this scenario, I should stress that I take H_t (and therefore, still, H_c) to be strictly conjectural and unproven.

Finally, in concluding this discussion of the Chinese Room argument, I should emphasise my admiration for the boldness of Searle's idea—that it might be possible to refute H_c *prior* to coming to any conclusion on H_t. Unfortunately, Searle's particular idea for doing this does not work.

2.4.2 Dualist Interactionism

It seems to me that, almost by definition, the only (realist) alternative to physicalism is some kind of pluralism; that is, one must suppose that there exist distinct classes of entity which interact with each other (they are, operationally, *real*) but which are not reducible to the class of physical entities (supposing, for the sake of the argument, that the latter class could be well defined in an unproblematic way). As far as mentality is concerned, this means a *dualist interactionist* position: holding that mental events are genuine entities, having causal effects on physical entities, but not themselves reducible to physical entities.

There is a distinction to be noted here between merely holding that physicalism is unproven (or even "unlikely"), and holding that it is actually false—i.e. *positively* advocating a dualist position.

Such a dualist position seems, however, not to be currently fashionable in the philosophy of mind. The *only* substantive contemporary example cited by Hofstadter and Dennett, in their extensive annotated bibliography of the field (Hofstadter & Dennett 1981, pp. 465–482), is that of Popper & Eccles (1977); I shall therefore give careful attention to a consideration of their position.

2.4.2.1 Criticism by Dennett

Dennett has provided a more or less detailed criticism of the position of Popper & Eccles, in the form of a book review (Dennett 1979). In Dennett's own words, this is a "caustic" review (Hofstadter & Dennett 1981, p. 477), where he finds very little of any sort to approve of, and appears to consider the arguments to be at best flawed, and at worst incoherent.

If Dennett were successful in his criticism, there would be nothing further for me to say here. However, while I generally agree with his conclusions, I consider that his route to them is quite inadequate, so there is still some work for me to do.

This inadequacy is presumably partly due to the constraints of the book review format. However, this cannot excuse, for example, Dennett's parenthetical summarising of Popper's *World 3* as "essentially a platonic world of abstract entities, such as theories, hypotheses, undiscovered mathematical theorems" (Dennett 1979, p. 94). The superficiality of this comment should be clear when it is noted that Popper actually expends several pages of argument to *distinguish* his World 3 from Plato's world of ideals (Popper & Eccles 1977, Chapter P2, Section 13).

Or again, Dennett severely criticises the apparent incompleteness of Popper's position:

> What kind of interaction can this be between a thinking and a theory? We are not told. Popper waves his hands at how modern physics has vacated all the old-fashioned philosophical ideas about causation, but does not give a positive account of this new kind of causation...
>
> Dennett (1979, p. 94)

But, when Eccles attempts to provide some analysis precisely of the nature of this causation, Dennett indulges his sarcasm from the opposite direction, accusing Eccles, in turn, of incompleteness because he:

> ... passes the buck to "the self-conscious mind," about whose apparently wonderful powers he is conveniently silent.
>
> Dennett (1979, p. 95)

Thus Dennett has managed to criticise each author for not covering issues dealt with by the other, and all this after peremptorily stating, in his introductory remarks, that:

> These men are not really co-authors, but co-contributors to an unedited anthology; they have not hammered out a joint theory, nor does it appear that they have been tough critics of each other's contributions.
>
> Dennett (1979, p. 92)

It seems that Dennett's review might have benefited from some tough criticism itself.

In summary then, while I agree with Dennett that the arguments propounded by Popper & Eccles are flawed, I consider that he has failed to confront them with the seriousness which they demand; and that, even where his criticism is well-founded, its credibility is undermined by embellishments which are not necessary, nor even consistent.

2.4.2.2 Eccles Neurophysiological Perspective

Eccles professes himself a dualist interactionist, but, as far as I have been able to establish, does not marshal any particular arguments in favour of this position. In his joint book with Popper, this issue is primarily dealt with in Chapter E7, where he expressly describes his purpose, not as the establishment of dualism as such, but as "the development of a new theory relating to *the manner* in which the self-conscious mind and the brain interact" (Popper & Eccles 1977, p. 355, emphasis added). That is, Eccles adopts the dualist interactionist hypothesis, *for whatever reasons,* and goes on to explore some of the consequences of this hypothesis; specifically, enquiring into the *nature* of the interaction between mind and brain.

I shall presume, though Eccles appears not to state it explicitly, that he relies on Popper for the prior establishment of the dualist position: his own rôle is then to consider some more specific implications of this general position. My task thus reduces to that of considering Popper's arguments alone; to the extent that I claim they are flawed, the considerations raised by Eccles are at least premature, if not irrelevant.[7]

2.4.2.3 Popper on AI

Popper is, at least, unambiguous in his view of what I have called H_c—he holds that it is false:

> I have said nothing so far about a question which has been debated quite a lot: whether we shall one day build a machine that can think. It has been much discussed under the title "Can Computers Think?". I would say without hesitation that they cannot, in spite of my unbounded respect for A.M. Turing who thought the opposite ... I predict that we shall not be able to build electronic computers with conscious subjective experience.
>
> Popper & Eccles (1977, Chapter P5, pp. 207–208)

Popper is less clear cut on H_t:

> Turing [Turing 1950] said something like this: specify the way in which you believe that a man is superior to a computer and I shall build a computer which refutes your belief. Turing's challenge should not be taken up; for any sufficiently precise specification could be used in principle to programme a computer. Also, the challenge was about behaviour—admittedly including verbal behaviour—rather than about subjective experience.
>
> Popper & Eccles (1977, Chapter P5, p. 208)

It seems that Popper accepts Turing's argument as showing that a suitably programmed computer may well be able to exhibit behaviour sufficient to pass the Turing Test (say); but considers *therefore* that there is little point in pursuing this. In particular, it will not necessarily endow a computer with "conscious subjective experience".

Thus far, Popper's position is quite comparable to that of Searle. However, his arguments for this position are entirely different, as we shall see.

[7] Eccles does make one other point that might be taken as a rationale for his dualist position—that he is "a believer in God and the supernatural" (Popper & Eccles 1977, p. VIII); but he does not expand any further on this, and thus there is no basis for substantive discussion here.

2.4.2.4 The Open Universe

Popper explicitly rejects physicalism, in all its manifestations, including what I have termed H_p. This is quite different from Searle who, as we saw, seems willing to accept the general idea of physicalism, rejecting only the special case represented by H_c.

Popper describes himself as a "dualist interactionist" with respect to the mind-body problem. However, he presents this in the context of his more general philosophy of the *Open Universe*, or what we might term a "pluralist" (rather than merely dualist) cosmology. That is, Popper holds that there exist, in the real universe, a variety of distinct classes of entities which are mutually interacting, but which are not reducible to each other; and that, furthermore, new irreducible classes of entity can, and do, *emerge* over time.

In particular, Popper has identified three specific classes of entities which, he claims, are not reducible to each other, and which he terms *Worlds*.

World 1 is the conventional world of unproblematic (?) physical entities. *World 2* is the world of subjective mental entities such as emotions, intentions, sensations, ideas, thoughts etc. Finally, *World 3* is the world of:

> ... products of the human mind, such as stories, explanatory myths, tools, scientific theories (whether true or false), scientific problems, social institutions, and works of art.
>
> Popper & Eccles (1977, Chapter P2, p. 38)

Thus, Popper specifically claims that World 1 and World 2 interact (they both contain *real* entities in good standing), but that they are mutually irreducible. This establishes his *dualist* position on the mind-body problem.

Popper has described the general idea of the Open Universe, and the Worlds 1, 2 and 3, in a wide variety of his writings. However, in what follows I shall restrict myself, for the most part, to the presentation of Popper & Eccles (1977), as this is where Popper explicitly relates this idea to the problem of artificial intelligence (or, at least, of artificial mentality).

Popper's attack on physicalism is two pronged: on the one hand, he identifies specific difficulties with a purely physicalist position; and on the other, he argues positively in favour of the dualist position. My rebuttal will therefore be similarly twofold.

2.4.2.5 Arguing Against Physicalism

Firstly, let me consider the specific difficulties alleged for physicalism. Popper provides a survey of varieties of physicalism, and adduces slightly different arguments against them. For my purposes, it is sufficient to concentrate on one specific variant, the *identity theory* (Popper & Eccles 1977, Chapter P3, Sections 22–23). Popper considers this the most difficult version of physicalism to rebut, going as far as to grant that, viewed in isolation, it *may* be true. However he claims that it is incompatible with Darwinism, and then argues that, since we must therefore choose between these two theories, we should prefer to retain Darwinism rather than physicalism.

My position is that Popper is mistaken in claiming that the identity theory (which is essentially equivalent to my H_p) is incompatible with Darwinism. Popper himself admits that his argument here is less than intuitively clear. It will require some care to deal properly with it—both to do justice to it in the first place, and then to answer it convincingly.

Popper's argument is that, under the identity theory, Darwinism is powerless to explain the *evolution* of mental entities, *per se*. This is so because:

- A Darwinian explanation can only work if the evolved entity has physical effects (roughly, it must positively affect the reproductive success of the carrier organisms).

- In the final analysis, under the identity theory, the mental entity can be shown to have physical effects *only* by replacing it with the (putative) physical entities with which it is identical.

- Such a purely physical Darwinian explanation, which has been shorn of all mental entities may, indeed, be valid. It will then properly explain why certain purely physical entities can evolve (i.e. because they are favoured by natural selection).

- However, since this explanatory scheme no longer contains any mental entities it is powerless to shed any light on why the (physical) entities which evolve are, in fact, identical with some mental entity.

- To put it another way, we would have a Darwinian explanation for the evolution of certain physical entities; we would *separately* know that these are identical to some mental entity; but this latter fact would have played no rôle in the evolutionary explanation. Thus, we could not then claim that the physical entities in question had evolved *by virtue* of this identity, nor of any properties of the mental entity, as such. We would have an explanation for the evolution of certain physical entities, but the fact that these are *also* correlated with (are identical to) some mental entities would stand as an independent, unexplained, and inexplicable, phenomenon. Indeed, according to our explanation they would have evolved in just the same way, even if they were *not* identical with some mental entity.

- That is, a Darwinian explanation for the specifically mental character of certain evolved physical entities is impossible. We would require some alternative explanatory principle, *in addition to Darwinism*, to address this.

- The incompatibility between the identity theory and Darwinism resides precisely in this result: that Darwinism would not be effective in explaining the evolution of mental entities.

I believe I have here stated Popper's argument in about as strong and as clear a form as is possible. I should add that Popper (Popper & Eccles 1977, Chapter P3, p. 88) also refers to a similar argument having been independently formulated by Beloff (1965).

I claim that the flaw in the argument is simply this: it goes through if and only if the characteristics of the physical entities which are relevant to their Darwinian selection are independent of (uncorrelated with) the characteristics which are relevant to their identification with some mental entity. To put it another way, an identification between a mental entity and some physical entities will, in the last analysis, require the physical entities to have some specific physical characteristics—otherwise the identification would be unwarranted. These physical characteristics may not be sufficient for the particular identification, but they would be necessary. Once this much is granted, it is unproblematic to incorporate these particular physical characteristics, which are essential elements of the identification, as factors in a Darwinian explanation of the evolution of the

(identified) mental entity.

To be specific, suppose that we have available to us a conjectural reduction of the entire mentality of some person to "unproblematic" physical entities: that is, we have a procedure for making identifications between the person's mental states and events and some physical states and events. A *necessary* (though not sufficient) condition for accepting this reduction, or system of identifications, is that the physical effects that result must be more or less consistent with the identified mental states and events—for example, the physical linguistic behaviour implied by the purely physical model must be consistent with the supposed mental states which correspond to it. To modify slightly an original example due to Fodor (1976, p. 199), one might postulate some particular identification which then turns out to have the property that a mental state of *believing that it will rain* predicts the consequent occurrence of the physical utterance "there aren't any aardvarks any more"; but one would then conclude that this identification between beliefs and physical states is, to say the least, suspect!

Ultimately, the core of Popper's argument seems to be this: if World 1 is causally closed (H_p is true), then Darwinism can, at best, provide an explanation of the evolution of certain physical phenomena, but these, in themselves, will have no *necessary* connection with subjective mental experience. Indeed, it seems to be apparent from Popper's criticism, already quoted, of the notion of Turing Testing, that he envisages that a system *could* well exhibit extremely complex behaviours, up to and including human level linguistic behaviours, and yet completely lack mentality; in a phrase commonly invoked by Harnad, it may be the case that, despite all appearances to the contrary, there could simply be "nobody home". If this is indeed possible—if the physical (including linguistic) manifestations of mentality can be had in the absence of mentality proper—then mentality would, from a Darwinian point of view, be redundant, and Darwinism would be incapable of explaining its evolution. But, if this *is* Popper's point, it seems to beg the question at issue: the idea of H_p (and, more specifically, of H_c) is precisely to conjecture that mentality proper—in the sense of "conscious subjective experience"—*is* a necessary correlate of certain physical behaviours. Now this conjecture may surely be mistaken, but it can hardly be criticised by an argument which already assumes it to be false.

The essence of the problem here for Popper, as previously for Searle, is to find an effective wedge to drive between H_c and H_t—for they both wish to accept the latter (tentatively, at least) but still reject the former. But once seen in this light, we can recognise that it is a very tall order indeed: it requires, more or less, a solution to the "other minds" problem—a basis for discriminating the mere "appearance" of mentality from "genuine" mentality. While Popper's approach is very different from Searle's, I cannot see that he is ultimately any more successful.

2.4.2.6 Arguing For Dualism

Next let us consider Popper's *positive* argument in favour of dualist interactionism (Popper 1973; Popper & Eccles 1977, Chapter P2).

The core of the argument is the claim that there exist at least some World 3 entities which are real (i.e. which interact, albeit indirectly, with World 1) but such that they are demonstrably *not* reducible to physical entities, i.e. are not *identifiable* with World 1 entities (they are "unembodied" in Popper's terms).

This would be enough to establish that the strictly physicalist view must be false. It would not, in itself, establish *mind-body* dualism, as such, i.e. the irreducibility of *World 2* to World 1. Popper completes the argument by pointing out that, in general, World 3 interacts with World 1 only through the mediation of World 2; therefore (so the argument goes), since World 3 itself is irreducible to World 1, and World 2 can interact with World 3, a capacity not exhibited by World 1 in general, then World 2 must *also* be irreducible to World 1.

I suggest that this latter argument is, in fact, defective. To see this, note that, under the identity theory (which Popper accepts "may" be true), the distinction between the mental and the physical is simply that certain states or organisations of World 1 entities do exhibit precisely the characteristics of World 2 entities, and, in this way, World 2 may be reduced to World 1. To apply this theory in Popper's scheme, we would simply stipulate that these distinguishing ("mental") characteristics of certain World 1 entities must include the ability to "grasp", as Popper puts it, World 3 entities. Popper has not offered any detailed theory of this interaction, which might show that it is beyond the ability of *some* such World 1 entities. Therefore, Popper has failed to justify the claim that interaction cannot happen *directly* between (unembodied) World 3 entities and (any) World 1

entities, and so has failed to establish the irreducibility of World 2 to World 1, as required for mind-body dualism.

This defect in Popper's argument is, indeed, pointed out by Dennett in his review Dennett (1979). However, his presentation is somewhat simplistic, if not actually mistaken—as when he says:

> It seems just as apt to say that when I put a Z brace on a gate to keep it from sagging, I bring about a causal interaction between theorems of Euclid and the pine boards, as it does to say that there is a causal interaction between my thinking and these theorems. That is, in the absence of much more detailed persuasions, both views appear ludicrous.
>
> Dennett (1979, p. 94)

It seems that Dennett is presenting here, not an argument as such, but an example of the kind of thing which he himself has described (in a different context) as an *intuition pump*—i.e. a thing which is "not, typically, an engine of discovery, but a persuader or pedagogical tool—a way of getting people to see things *your* way once you've seen the truth" (Dennett 1980).

I suggest that, in fact, the attempted *reductio ad absurdum* fails to fully confront Popper's argument. For the *crux* of Popper's argument is not that World 3 entities, *in general*, can only interact with World 1 via World 2 (though he does, admittedly, claim this); the important point is the much more particular claim that this is so for certain *specific* World 3 entities, namely those which are "unembodied" or demonstrably irreducible to World 1 entities. For Dennett to properly refute Popper's argument, his example of a direct interaction between World 3 and World 1 would have had to involve some such *unembodied* World 3 entity, rather than just any arbitrary World 3 entity. His example is not of this sort; or at least, is not *clearly* so. I shall return to this below.

Thus, while I have restated Dennett's point—that Popper has failed to establish that unembodied World 3 entities *cannot* interact directly with World 1—I have not relied for this on Dennett's suggestion that unembodied World 3 entities can, in fact, interact directly with simple unproblematic World 1 entities such as pine boards. Rather, I am willing to stipulate, with Popper, that the interaction requires the mediation of World 2; but then point out that this observation, in itself, is neutral with respect to the reducibility of World 2, *in some fashion*, to World 1.

The flaw in Popper's argument is, then, that he (implicitly) proceeds from the premise that *certain* unembodied World 3 entities cannot interact directly with *certain* World 1 entities (this would include, for example, Dennett's pine gate), to the conclusion that unembodied World 3 entities cannot interact directly with *any* World 1 entities (such as minds, or rather, under the identity theory, the putative World 1 entities which are identifiable with minds). In taking this step he *assumes* the irreducibility of World 2 (i.e. the non-existence of World 1 entities which are identifiable with minds), which is precisely what he is purporting to establish. In short, his argument fails because it is ultimately circular.

However, whether one accepts Dennett's simplified analysis, or insists upon the more detailed refutation presented here, the outcome is actually still peculiarly unsatisfying.

We see that Popper's conclusion of mind-body dualism is unwarranted, because one particular step in his argument is defective. From both Dennett's point of view and my own, this is, arguably *enough*: we have provided a sufficient basis to refute Popper's argument for mind-body dualism, which is all we really sought to do; and, indeed, that is where Dennett does leave the issue. But: it involves attacking Popper on the *weakest* element of his argument, while still leaving his central, substantive, point unchallenged.

This central point is the claim that World 1 is causally *open*—that there exist entities which are demonstrably not reducible to World 1 entities, but which are perfectly real in the sense of *altering* the behaviours of some World 1 entities from what would be predicted based solely on their interactions with the rest of World 1.

It would be much more satisfactory if one could sustain a challenge against Popper's argument for an Open Universe as such, rather than relying on a rather technical nicety in how he has applied it to the issue of mind-body dualism. This is precisely what I shall now try to do.

The critical step is Popper's claim that certain World 3 entities are "unembodied", i.e. irreducible to World 1 (or World 2, for that matter), but, nonetheless, have definite causal effects on World 1 (via World 2).

The first part of this is unobjectionable: Popper is the originator of the World 3 concept, so he is surely entitled to include within it whatever he wishes.

In particular, he may include things like *unproved theorems*: that is, statements which are, in fact, true (relative to some system of axioms) but for which no one has yet actually found a proof. By definition, such things are, indeed, unembodied—there do not exist any World 1 or World 2 entities correlated with them.

It is the second part of Popper's claim that seems to me to be potentially problematic: the assertion that such unembodied World 3 entities are *real,* in the sense of interacting directly with World 2, and thus indirectly (at least) with World 1. Popper deals explicitly with this issue as follows:

> ... Thus a not yet discovered and not yet embodied logical problem situation may prove decisive for our thought processes, and may lead to actions with repercussions in the physical World 1, for example to a publication. (An example would be the search for, and the discovery of, a suspected new proof of a mathematical theorem.)
>
> Popper & Eccles (1977, Chapter P2, p. 46)

If I understand him correctly, Popper's point here is that the truth of a mathematical theorem (for example) is an objective World 3 fact which is independent of any embodiment in World 2; it is, indeed, as objective as any World 1 fact. In particular, it is intersubjectively testable. Such tests are always fallible of course—but so too are tests of supposed World 1 "facts". Since these World 3 facts can exist and persist despite not being embodied, they evidently (?) cannot be reduced, without residue, to World 2 or World 1 entities; but since they *can* interact with World 2 (or be "grasped"), and thus with World 1, they are surely *real.* Popper's conclusion is then that World 1 cannot be causally closed.

This is a highly original and bold argument. It is, intuitively, quite compelling. And yet, when I examine it critically, it seems to me that it has very little substance, and cannot possibly be made to bear the burden which Popper attempts to place upon it.

Let us consider Popper's own favoured example: the truth of a mathematical theorem. This objective World 3 entity may be said to "interact" with a mathematician in the sense of constraining her results; she will not, in particular, be able to prove the theorem, nor any of its corollaries, *false*, no matter how hard she may try; the reality of the theorem may be said to manifest itself through the failure of such attempts. This is so, regardless of whether the mathematician

ever explicitly conjectures, even, that this theorem exists. Let me stipulate, then, that this establishes the "reality" of the theorem.

The *irreducibility* of the theorem is separately held to follow from the fact that, at a given time, there may be nobody at all (no World 2 entities) who have yet even conjectured that it may hold, so there are not even any *candidate* World 2 entities as targets for a reduction (and thus, surely, there are no World 1 candidates either). But this claim is just wrong.

The theorem, if it *is* a theorem, is already implicit in the axioms of the system under study; it may be said to exist at all (in Popper's sense) only when some such axioms have been already *adopted.* That being the case, there is a perfectly good sense in which the theorem may be "reduced" to the *axioms*; and (by hypothesis) the axioms *are* already embodied, and thus *are* potentially reducible to World 2 (and ultimately even World 1) entities.

The point can be made more definite by replacing Popper's mathematician by a theorem proving *machine.* Such machines have indeed been built. By Popper's own hypothesis, such machines lack mentality, so they are not World 2 objects. Yet they can interact with, be constrained by, or even "grasp", the truth of a theorem in precisely the sense outlined above for a (human) mathematician. And they do so simply because this World 3 object, this truth of a theorem, is no more and no less than a product of the inference rules with which the machine was originally equipped. But the system in question here is a paradigm example of a causally closed physical (World 1) system. While it is true that, initially, the machine has no explicit embodiment of the theorem (even as a conjecture), this plainly does *not* establish (*pace* Popper) that the theorem is irreducible to World 1, or that the machine, *qua* World 1 entity, must be causally open to some influences which are not in World 1.

This example may be said to refine and elaborate the earlier example suggested by Dennett, of an interaction between theorems of Euclid and a Z brace. It goes beyond Dennett's example just insofar as it stipulates that, in the initial state of the physical system, the theorem in question is only implicit in a set of axioms—and is thus "unembodied" in Popper's sense—and yet "interacts" with the system in just the sense with which it might be said to interact with a mathematician.

However, I am not sure that this quite exhausts Popper's argument yet. Popper is well aware of the possibility of theorem proving machines (though I am not aware of his having analysed their implications in just the way I have suggested above). Thus, even before he had fully formulated the concept of World 3, he made the following remark (this originally dates from c. 1957):

> A calculator may be able to turn out mathematical theorems. It may distinguish proofs from non-proofs—and thereby certain theorems from non-theorems. But it will not distinguish difficult and ingenious proofs and interesting theorems from dull and uninteresting ones. It will thus 'know' too much—far too much—that is without any interest. The knowledge of a calculator, however systematic, is like a sea of truisms in which a few particles of gold—of valuable information—may be suspended. (Catching these particles may be as difficult, and more boring, than trying to get them without a calculator.) It is only man, with his problems, who can lend significance to the calculators' senseless power of producing truths.
>
> Popper (1988, pp. 107–108)

This suggests to me a different, and more nebulous, interpretation of Popper's ideas. While I believe that the existence of theorem proving machines (even those proving uninteresting theorems!) adequately rebuts Popper's later, specific, claim that "unembodied" theorems are necessarily irreducible to World 2 or World 1, it seems that Popper might not wish to rely on that argument anyway—that he has a much more general notion of an irreducible World 3 in mind. This is borne out, to an extent, in the following comment:

> There is no doubt in my mind that the worlds 2 and 3 do interact. If we try to grasp or understand a theory, or to remember a symphony, then our minds are causally influenced; not merely by a brain-stored memory of noises, but at least in part by the autonomous inner structures of the world 3 objects which we try to grasp.
>
> Popper (1973, p. 25)

To return again to the mathematician, it seems that Popper may wish to claim something much stronger than anything I have so far discussed. He may conceivably mean something like the following: that the objective existence of a theorem may change the pattern of the mathematician's thoughts so that (for example) she moves towards its formulation (or proof), in a way that is *not* already implied by her prior thoughts–i.e. in a way above and beyond the explanatory power of purely World 2 entities (noting of course, that the relevant World 2 entities will presumably be embodying certain World 3 entities). This should be

contrasted sharply with a claim merely that the mathematician's *suspicions* or *intuitions* about the theorem affected her thought processes (as they undoubtedly would); for suspicions and intuitions are common or garden World 2 objects (presumably correlated with World 3 entities—but, by definition then, *these* are *already* embodied).

But it should be clear that any such interaction between unembodied World 3 entities and World 2 must be, at best, conjectural—one possible interpretation of the example of the mathematician, but not at all a conclusion from it. Indeed, if we apply Popper's own criteria for the evaluation of scientific theories, we should say that the hypothesis that unembodied World 3 entities do *not* have such causal effects on World 2 has a greater *content* (and thus corroborability) than its converse, and, in the absence of some evidence that it has actually been refuted (and none is offered, that I can see) should be preferred, even if only for the time being.

But the ramifications run deeper: such interactions between World 3 and World 2 would be completely inconsistent with the rest of Popper's evolutionary epistemology. They would be tantamount to a form of Lamarckian instruction by World 3 of World 2—i.e. Lamarckism applied to the evolutionary growth of an individual's subjective knowledge. This is something that has been resolutely opposed by Popper in the case of knowledge of World 1 (he has dubbed it the "bucket" theory of knowledge—Popper 1949; 1970b), and I see no reason why his arguments should have any less force in the case of our knowledge of World 3. I therefore conclude that this cannot, after all, be a plausible interpretation of Popper's position.

It is important to note that none of my discussion here attempts to deny the reality of World 3 (an attack anticipated by Popper). I claim only that Popper has not established the irreducibility of World 3 to World 2 (and thus, possibly even to World 1). World 3 is still a perfectly meaningful and useful idea; as long as we admit that its reducibility is an open question, and that the hypothesis that it *is* reducible is actually stronger (has greater content) than the converse, and is currently a preferable basis for research.

2.5 Conclusion

In summary, my claims in this chapter are:

1. That physicalism in general, and computationalism in particular, are irredeemably repugnant to human values and to the dignity of mankind; it seems to me craven to deny this.

2. That neither physicalism nor computationalism have (yet) been definitively refuted; in particular, two distinct kinds of argument, by Searle and Popper respectively, purporting to achieve such a refutation, are flawed.

The relationship between these two points is, I think, very important. It seems to me that the first point precisely underlies the *intuitive* conviction of those, like Popper and Searle, who hold that H_c is definitely false. It should be clear that I completely share this intuitive conviction; I will confess, if that is the correct word, to being a *metaphysical* dualist.

However: the point at issue is how we might proceed *beyond* intuition. This raises what is almost a refrain of Popper himself:

> I regard intuition and imagination as immensely important: we need them to invent a theory. But intuition, just because it may persuade and convince us of the truth of what we have intuited, may badly mislead us: it is an invaluable helper, but also a dangerous helper, for it tends to make us uncritical. We must always meet it with respect, with gratitude, and with an effort to be severely critical of it.
>
> Popper (1988, Preface 1982, p. xxii)

Both Popper and Searle have attempted to proceed by supporting their intuitions with definite arguments—arguments which come close to having a scientific rather than a metaphysical character. If these arguments were acceptable—if H_c, in particular, were thereby refuted—then further investigations within the computationalist framework (such as, for example, attempts to *realise* Turing Test capability with computational systems) could only have technological significance; such investigations, though potentially valuable in their own right, would no longer directly bear on what, in Chapter 1, I called *Popper's Problem*—the cosmological problem of understanding the world and our place in it. Thus, if one wished to remain focused on this latter problem then one would be led, instead,

to proceed with a programme of research which reflected and incorporated the refutation of computationalism. Such an approach might be typified by the work of Eccles on the "liaison" between mind and brain, for example.

But I have claimed that the arguments put forward by Searle and Popper are flawed, and do not support the conclusions claimed. In particular, while I remain intuitively convinced of the falsity of H_c, this remains, for me, a *merely* intuitive belief. So the question remains of how best to proceed. Somewhat ironically, I think Popper has already suggested at least one possible answer to this:

> ... as a philosopher who looks at this world of ours, with us in it, I indeed dispair of any ultimate reduction. But as a methodologist this does not lead me to an anti-reductionist research programme. It only leads to the prediction that with the growth of our attempted reductions, our knowledge, and our universe of unsolved problems, will expand.
>
> Popper (1974c, p. 277)

The programme of computationalism—of attempting to realise or synthesise the "appearances" (at least) of mentality by computational means—is an essentially reductionist one. Like Popper, I too do not expect any kind of ultimate success from this effort. But our failures, and the precise mechanisms of these failures, may be extremely interesting, and perhaps even revealing. There is thus every reason to pursue this programme of "methodological computationalism", despite our pessimism about its potential for "success"—just so long as we can avoid dogmatism, and continue to be critical of it. For the remainder of the Thesis then, I shall tentatively *adopt* this computationalist thesis, H_c, and explore at least some of its detailed ramifications.

Chapter 3

Artificial Knowledge

3.1 Introduction

This chapter moves on from the metaphysical consideration of what kind of a thing a mind is, or might be, to the pragmatic consideration of building machines (especially computers) that exhibit some or all of the *behaviours* associated with mentality—which is to say, a consideration of Artificial Intelligence (AI) in what Searle (1980) calls the "weak" sense. Alternatively this may be viewed as an investigation of the hypothesis I have previously (Chapter 2, section 2.2) called *Turing Test Computationalism* (H_t)—the claim that a suitably programmed universal computer could pass the Turing Test.

I start with a brief review of the Turing Test itself, and, in particular, some novel criticisms of it proposed by French (1990). I shall consider these criticisms, but argue that the Test still stands as a valuable focus for work in AI; nonetheless, I shall go on to conclude that performance at this level is still so far beyond our present theoretical understanding, that Turing Testing, as such, may of little immediate or practical interest.

I next consider the general issue of *cognitive architecture*—what, if anything, can we say about the overall structure which a (computational) system must have if it is to exhibit behaviours indicative of intelligence. The essential point I make is the negative one that *universality* (in the technical sense characterised by, for example, Universal Turing Machines), *per se,* does not mean that a computational

intelligence will admit of explanation in terms of a unitary "symbol level", or "language of thought".

I then consider the notions of "meaning" and "knowledge" in more detail, in an effort to show that a *computational semantics* is indeed possible (despite some claims to the contrary), and I sketch out what it might look like. In particular, I claim that computers can realise *anticipatory systems* (Rosen 1985a), and that, in this case, they exhibit *intentionality* (Dennett 1978b), and instantiate *subjective knowledge* in the sense that Popper admits for biological organisms generally (e.g. Popper 1961). These claims are made *independently* of any commitment to the idea that computers are able to realise "genuine" mentality—in the sense of "conscious subjective experience".

With this particular philosophical perspective, I then briefly consider methodological approaches to AI, in particular the notion of "Knowledge Engineering". I note that this approach has run into serious difficulties, typically identified with the *common-sense knowledge* problem. It has proven extremely difficult to *explicitly* formulate common-sense knowledge (and thus incorporate it into computer systems). There is little general agreement as to the nature of this problem; but it seems that developing an explicit, brute force, *stipulation* or *enumeration* of common-sense knowledge is currently an intractable problem, and may yet prove to be completely impossible.

The alternative to the Knowledge Engineering approach is, of course, to develop some kind of "adaptive" or "learning" system; which is to say, we turn from the problem of knowledge in itself, to the rather different problem of its *growth.*

I shall argue, from several different points of view, but based particularly on the *evolutionary epistemology* pioneered by Popper and D.T. Campbell, that a kind of abstract generalisation of *Darwinian* processes, referred to as *Unjustified Variation and Selective Retention* (UVSR), is an essential component in the growth of knowledge. I conclude from this that the realisation of *Artificial Darwinism* may a necessary, though certainly not sufficient, condition for the realisation of Artificial Intelligence.

3.2 The Turing Test

3.2.1 Definition

In his influential paper, *Computing Machinery and Intelligence* (Turing 1950), Alan Turing set out to consider the question "Can machines think?" (p. 433); ultimately, however, he concluded that, in this form, the question was "too meaningless to deserve discussion" (p. 442). Instead, Turing proposed an *operational* definition for "thinking", and restricted "machine" to designate a suitably programmed *digital computer*. He then considered the new question of whether such a machine could satisfy such a definition.

This operational definition of thinking was phrased in terms of what Turing called the "Imitation Game", and is now generally referred to as the *Turing Test*.

Briefly, the Turing Test involves a human *interrogator*, and two *subjects*. One subject is the *machine* to be tested, the other is a human *control*. The interrogator is restricted to interacting with the two subjects purely linguistically (for example, via teletype), and has no other way of distinguishing between them. One *turn* then consists of a fixed time—Turing suggests 5 minutes—in which the interrogator is allowed to question the subjects, after which time he must nominate which subject he judges to be the human and which the machine. The machine is considered to have *passed* the Test if the interrogator's *probability* of making a successful classification is found to be below some specified threshold—Turing suggests 70%. Turing omitted various details here: one presumes that the success probability would be measured by playing out as many turns as are necessary to get a statistically significant result, while varying the interrogators and control subjects to achieve independence between turns. Turing does explicitly refer to the use of an "average" interrogator (p. 442).

3.2.2 Sufficiency?

As discussed in Chapter 2, there is room for argument as to the *sufficiency* of the Turing Test. That is, whether an entity's ability to pass this Test is a sufficient condition for saying that it exhibits *mentality*. If the Test were not sufficient in this sense then that would certainly limit its interest. However I have already

stated, in Chapter 2, my opinion that the proposed arguments to such an effect are far from compelling; and that I shall therefore proceed on the basis that, *pro tem,* Turing Testing *is* a sufficient operational criterion for mentality.

3.2.3 Necessity?

French (1990) takes the position that the Turing Test is valid or sufficient for the attribution of intelligence, but argues that it is in fact much *more* stringent than Turing anticipated or intended. Specifically, he suggests that "... the Turing Test could be passed only by things that have experienced the world as we have experienced it" (p. 53). While he believes that *in principle* a machine could indeed be built which would satisfy this constraint, he assumes that, *in practice,* "no computer is now, or will in the foreseeable future be, in a position to do so" (p. 56). It follows, of course, that there is no practical prospect of a computer passing the Turing Test. French concludes that some alternative tests, or at least criteria, are therefore needed for practical use in AI research.

I should emphasise that I agree with French on certain points which he raises. For example, he suggests that the Turing Test is deficient in that it admits of no "degrees" of intelligence, and is not applicable at all to non-linguistic behaviour that might, in fact, be related to intelligence (such as exhibited by animals). I agree with this as far as it goes: given that Turing Test performance is currently an intractable problem, it is sensible to formulate lesser or entirely distinct criteria which might, once achieved, represent progress toward that ultimate goal. In fact, this is what goes on in practical AI research all the time.

Where I disagree with French is when he goes on to suggest that the Turing Test should be dispensed with altogether, even as an *ultimate goal* against which intermediate goals can and should be critically reviewed. Even here, I shall give some ground, though not, I think, as much as French seeks.

French's argument is that the Test, as formulated by Turing, admits the use of so-called *subcognitive probing* by the interrogator, and that this makes the procedure an unnecessarily *harsh* or severe test of general intelligence. That is, French supposes that there could be systems (presumably including certain suitably programmed computers?) which would be unable to pass the Turing Test,

but which should, nonetheless, be labelled intelligent—indeed, "as" intelligent as humans, if not more so.

The idea of subcognitive probing is to ask questions which, in some sense, probe the underlying, subconscious, "structure" (the *associative concept network*[1]) of the putatively intelligent subject. French argues that this is possible under the Turing Test conditions, and that it would allow specifically or uniquely human aspects of intelligence to be detected—aspects which would be very difficult, if not entirely impractical, to duplicate in a computer, and which are, in any case, *inessential to general intelligence.*

In fact, French concludes that the practical development of some entity such that it could pass the Turing Test, given the use of subcognitive probing, would require that the entity be capable of experiencing the world "in a manner indistinguishable from a human being—a machine that can fall off bicycles, be scratched by thorns on roses, smell sewage, and taste strawberries..." (French 1990, p. 56): that is, the system would have to be a more or less humanoid robot or *android.* It is this scenario which French regards as being impractical (though not, in principle, impossible) for the foreseeable future. More to the point, he considers that this renders the Turing Test unsuitable for practical use.

French further claims that the Turing Test cannot be modified, in any reasonable way, so as to eliminate the possibility of subcognitive probing, and should therefore be simply discarded. He does not propose a specific, operational, alternative, but suggests that we should consider intelligence in the "more elusive terms of the ability to categorise, to generalize, to make analogies, to learn, and so on" (p. 65).

I agree with French that the use of subcognitive probing, as he describes it, would subvert the Turing Test; that, indeed, such probing is one of the general kinds of thing Turing was trying to preempt in his design of the Test; and that only some test which does not exploit such probing would be satisfactory. However, I disagree with French that such probing cannot be eliminated from the Turing Test, with little or no modification. I shall argue this on several grounds.

[1] French presumes that some such network necessarily underlies intelligence; I do not disagree as such, but it might have been better if he had made his assumption explicit, and phrased it as an *hypothesis,* rather than taking it as some kind of established fact.

First, and most obviously, French is able to introduce subcognitive probing in the first place only by effectively changing (or, at least, augmenting) the rules of the original Test. Specifically he requires that the interrogator be allowed to *poll* humans for the answers to some questions prior to posing them during the Test itself. This is in order to allow statistical analysis of the "subcognitive" characteristics of responses to these questions, as exhibited by people, so that these could then be compared with the behaviours of the subjects in the Test proper. French states that he feels "certain" that Turing would have accepted this. I happen to disagree with this opinion, but it is irrelevant in any case. The point is that if we *disallow* such polling (whether Turing would have approved or not) the Test is effectively immunised against the use of subcognitive probing, by French's own admission.

But quite aside from this, I think French's analysis is contrived and mistaken. While Turing did not specify precisely what he meant by an "average" interrogator, it seems absurd to suppose that he would have allowed interrogators who are familiar with, and competent to consciously apply, the notion of subcognitive probing. Again, of course, the question of what Turing's own opinion might have been is strictly irrelevant anyway: the important point is that, in response to French's criticism, we are quite free to add an explicit stipulation to the Test, to the effect that persons having a competence in the technique of subcognitive probing will not be allowed as interrogators—*if* that is deemed necessary in order to eliminate subcognitive probing. In fact, I suggest that, for virtually any practical purposes, it would be adequate simply to *stipulate* to interrogators, at the start of any Test, that they must not attempt to *use* subcognitive probing in their evaluation of the subjects.

French might still argue for the possibility of *unconscious* subcognitive probing having some statistically significant effect on the Test outcome. This would obviously be, at best, a much weaker argument, and I don't believe it could be sustained in any case. Remember that the Test, as Turing specified it, is relatively coarse (presumably deliberately?): the interrogators' success rate only has to fall below about 70% for the computer to pass. I doubt very much that a credible argument could be made to the effect that subcognitive factors, *alone,* are likely

to consistently, and unconsciously, bias the success rate by 30 percentage points or more.

But even if, despite its intuitive implausibility, we suppose that French could marshal enough evidence to show an effect of this magnitude due solely to unconscious subcognitive factors, I claim that the effect could *still* be nullified with relative ease. This can be done by interposing what I shall call a *subcognitive scrambler* between the interrogator and the subjects. This would simply be another person, who relays all messages between the interrogator and the subjects. The interrogator is now restricted to have direct access only to the scrambler, and not to the subjects. The scrambler takes up the previous position of the interrogator, having linguistic access to the subjects, via teletype or otherwise, but otherwise having no knowledge of the identities of the subjects. The sole instruction to the scrambler is to *paraphrase* the messages passed from interrogator to subjects, and back, in such a way as to maintain their essential semantic content, but to otherwise modify them as much as he wishes. A particularly effective way to achieve this might be to use interrogators whose native language is different from that of the subjects, and thus have a *translator* act as the subcognitive scrambler.[2]

I freely admit that such a scrambler would not be effective against *all* kinds of deliberate or conscious attempts at subcognitive probing.[3] However, I think it would greatly attenuate any possible *subconscious* subcognitive effects, which was the remaining point at issue.

In conclusion then, I consider that the deficiency in the Turing Test, alleged by French (i.e. its supposedly *excessive* stringency), is either non-existent or easily corrected, and the Test can therefore survive his attack more or less unscathed.

---

[2]In allowing, or even recommending, the use of such translation, I implicitly transgress, to at least some extent, against another assumption which French allowed himself: that the human subject and the interrogator "are all from the same culture and that the computer will be attempting to pass as an individual from that culture". Again, I see this as *ad hoc* and contrived on French's part, and not sustainable.

[3]I have in mind specifically what French calls the *Category Rating Game* technique.

3.2.4 An Informal Test

And yet: while I disagree with French's literal arguments, I cannot help but believe that there is some core of truth about his ideas.

Let me suggest then that detailed, legalistic, discussion of the Turing Test is pedantic, and essentially futile—notwithstanding the fact that I have just indulged in such a discussion above. I indulged in it because that is the ground on which French had chosen to mount his assault, so I wished to respond in kind; demonstrating that, judged even in his own terms, his assault founders. However, in many ways it was a pity that Turing gave a relatively precise description of his proposed Test—for it is this spurious precision that prompts excessive concentration on the details, such as exhibited by French.

I suggest that the Turing Test should best be considered as a blunt (though moderately effective) instrument, whose details are entirely unimportant. Its point lies not in any detailed experimental set up, but in the *principle* that any machine which can credibly, or meaningfully, participate in human conversation should, *regardless of what other attributes it may have* (especially its physical constitution), be regarded as a *bona fide* member of the community of sentient beings.

I suggest especially that *indistinguishability* between machine and human conversation, which is at the core of much discussion of the Test, including that of French, is actually a red herring. I think that this is implicit in the rather coarse tolerance of 70% originally suggested by Turing for his Test.

The real issue is *credibility*: whether some putative machine intelligence can sustain a conversation in such a way that we would be satisfied that it really *means* what it says; this remains the case, even if what it is saying is obviously and thoroughly non-human (and thus perfectly "distinguishable" from human conversation). For example, the conversation that would be involved in actually inviting a machine to act as a subject in a formal Turing Test would certainly involve elements that would not arise in any normal conversation between human beings; but I suspect that, on the basis of just such a conversation, one could sensibly judge whether the machine *meant* what it was saying or not.

So, if French's point is that the Turing Test, as stated, focuses on indistin-

guishability from strictly human intelligence, and that this is unnecessary and even misguided, then I am inclined to agree with him. French however, sees this as an *intrinsic* defect of the Test. I think he is mistaken in this, as I have already argued; but even if he were right, I think this conclusion would be contingent on a very literal reading of the Test (which, I admit, overemphasises the issue of comparing machine with human intelligence), and a consequent failure to appreciate the central, informal, idea being promoted by Turing.

What I take to be the proper view of the Turing Test has been previously elaborated by Roger Penrose:

> From my own point of view I should be prepared to weaken the requirements of the Turing test very considerably. It seems to me that asking the computer to imitate a human being so closely so (*sic*) as to be indistinguishable from one in the relevant ways is really asking more of the computer than necessary. All I would myself ask for would be that our perceptive interrogator should really feel convinced, from the nature of the computer's replies, that there is a *conscious presence* underlying these replies—albeit a possibly alien one. This is something manifestly absent from all computer systems that have been constructed to date.
>
> Penrose (1990, p. 11, original emphasis)

To be clear then, let me now propose what I shall dub the *Penrose Test* for intelligence:[4]

> Any entity is intelligent which understands and means what it says; and any entity understands and means what is says if it can so convince a competent human judge, purely on the basis of her conversation with it.

By a "competent" judge I mean someone who, *inter alia,* has a reasonable understanding of the state of the art in AI, and is capable thereby of probing past "canned" responses of the so-called ELIZA type (Weizenbaum 1984), etc. Indeed the judge should probably have some specific familiarity with whatever design principles may have been used in building the putative intelligence (in this limited respect, the test I propose here is arguably *more* stringent than Turing's).

[4]Perhaps this might equally be called the *Asimov Test* for intelligence; compare it with this formulation: "There is no right to deny freedom to any object with a mind advanced enough to grasp the concept and desire the state" (Asimov 1976, p. 174). The "grasping" and "desiring" are apparently to be established by similar criteria to those which I have suggested: linguistic cross-examination of the subject. In Asimov's case a "competent" judge is, effectively, any court of law having relevant jurisdiction.

I have omitted any comment about the allowed or required *domain of discourse* in the Penrose Test. This is deliberate. I consider that the demand that the entity convince a competent judge, purely through conversation, that it really does understand and mean what it says is already enough to guarantee a satisfactorily wide ranging domain of discourse, without any additional stipulation.

My claim is that the Penrose Test captures the essence of Turing's original Test; and that, in particular, any honest researcher can judge perfectly well whether his system should be labelled intelligent, in this sense, without ever having recourse to the elaborate paraphernalia actually prescribed by Turing, and without any significant danger of being confounded by irrelevant factors, subcognitive or otherwise.

Furthermore, I suggest that this is in fact the way the "Turing Test" is employed by practical researchers. I think it is generally accepted that no AI system yet developed has come remotely close to meeting Turing's criterion, and this is known without any attempts at setting up the kind of formal test conditions actually described by Turing. The latter would only come into play if or when we have a system which we *already* know, from the Penrose Test, to have the depth of understanding required to participate in a meaningful conversation; but even then the formal Turing Test would, at best, serve only to demonstrate the objectivity of this claim. And of course, we should remember that the rôle of the machine in the Turing Test is distinctly demeaning, if not positively insulting: it seems to me that a *prima facie* mind might well refuse to participate in such a charade!

For a more detailed discussion of the issues arising here, see Hofstadter's (1985, Chapter 22, Post Scriptum) account of actually attempting to apply Turing's ideas on testing intelligence *in practice* (albeit the "intelligence" being tested turned out to be a hoax—a gentle practical joke at Hofstadter's expense). I consider it significant that, in operation, this turned out to be much closer to my description of the *Penrose* Test than a *Turing* Test proper. Hofstadter also, incidentally, anticipates the notion of subcognitive probing, subsequently elaborated by French. Notwithstanding this, Hofstadter's conclusion, at that time

at least, was that he was (still) an "unabashed pusher of the Turing Test as a way of operationally defining what it would be for a machine to genuinely think" (p. 525).

So my final answer to French is to strictly disagree with his criticism, and insist that the Turing Test is still essentially as satisfactory as when Turing first proposed it; but I admit that the formal aspects of the Test are distracting, and I actually propose the Penrose Test as a clarification of Turing's central idea. Indeed, while I shall continue to refer to "Turing" testing in what follows, this should now be interpreted (where this makes any difference) as *Penrose* testing.

3.3 The Problem Situation in AI

Turing's answer to his own reformulated version of the question of machine intelligence was that he believed that a suitably programmed digital computer probably *could* pass his Test. Indeed, he went so far as to predict that "at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted" (Turing 1950, p. 442). This was perhaps somewhat rash. It is now clear that the implied target of programming a computer such that it is capable of passing the Turing Test, by the end of the century, will not be achieved; indeed, there is little consensus as to when, *if ever,* a Test might (with any confidence) be rescheduled for![5]

To be fair to Turing, he was not at all dogmatic. He explicitly stipulated that his claim (that a computer could be made to pass the Test) was conjectural and speculative; that he had no decisive arguments to show that it was possible (even in principle—never mind in practice); and that its truth could only be definitively established by exhibiting a working example of such an "intelligent" computer. Of course, it *was* an essential part of Turing's paper to consider and

[5]Granted, the annual *Loebner Prize Competition*, launched in 1991, is derived from the idea of the Turing Test (Campbell & Fejer 1991). However, it is based on an extremely impoverished version of the Test, in that each subject can only be interrogated on a single, specified, topic, and the interrogators "were told to hold normal conversations, not to try aggressively to unmask the contestants with tricky questions" (Strok 1991). I note that the 1991 prize for the best performing computer subject (whose topic was "whimsical conversation"!) was presented, with no apparent sense of irony, not to the subject itself but to its programmer.

discount arguments *against* even the possibility of a computer passing the Test: for otherwise the formulation of the Test would have been pointless. By way of conclusion, Turing admitted that, at the time of writing, it was very unclear how best to go about trying to make a computer pass the Test, or even what the basic hardware requirements might be. Thus, Turing's achievement was in sharply defining an interesting problem, rather than offering a substantive theoretical insight into its solution. It is my view that the problem situation in Artificial Intelligence can still be quite well characterised in the way outlined by Turing. Specifically, I suggest that:

- The Turing Test has not been shown to be *invalid* (i.e. not a sufficient test for intelligence). Indeed, I would argue that this question will not actually become pressing until (or unless) some system other than a human being (whether a programmed computer, or something else, as yet unimagined) actually passes it.

- Turing Test performance has not been shown to be impossible, or inherently impractical, for a computer (not even if we restrict attention to those computers which are *already* technically feasible).

- Conversely, no essentially new argument has been forthcoming to suggest that Turing Test performance definitely *is* possible (even in principle) for a computer (either now or in the future).

- We still lack any comprehensive understanding (theory) of what, specifically, would be required to make a computer pass the test. There has, of course, been a major research effort over the 40 years since Turing's original assessment of the situation. This has yielded considerable insights into the problem. I review some of this work in subsequent sections. There is no doubt that our understanding of the *difficulties* in achieving Turing Test performance from a computer is now much more acute than when Turing first formulated the problem; but it is certainly *not* the case that we now know "in principle" how to achieve this performance, but only lack (say) adequate hardware, or a sufficient software development effort, to bring it about.

3.4 On Cognitive Architecture

Turing, through the notion of the Universal Computer, provided an *existential* argument for Artificial Intelligence: it seems that *some* (universal) computer, running *some* program, should "surely" be able to pass the Turing Test. This is the hypothesis of *Turing Test Computationalism* (H_t).

Turing Test performance would mean (by definition) that we impute "mental" states and events to such a machine. This can be done even without a commitment to the view that the machine "really" has any mentality: Dennett (1971) refers to this process as the adoption of the *intentional stance*. More generally, even for a machine which does not achieve full Turing Test performance, the behaviour may still be such as to justify a limited adoption of the intentional stance, i.e. the imputation of mentality, albeit in some impoverished form. In this way we sidestep, even if only *pro tem*, the metaphysical debate as what "genuine" (human) mentality actually consists in.[6]

Since the *un*-programmed computer manifestly lacks mentality, we thus effectively impute mentality to the computer *program(s)*: that is, the general notion that there can exist programmed computers which are intentional, or the more particular notion that there can exist programmed computers which can pass the Turing Test (H_t), implicitly asserts that the mental states and events of (or imputed to) such machines are, in principle, reducible to, or identifiable with, states and events of their programs, which is to say of purely "computational" entities.

The point is that, *whenever* we adopt the intentional stance toward a programmed computer, we implicitly identify some reduction of mental entities to computational entities.

It is the *nature* of these reductions that is actually of central interest—particularly, though not exclusively, for whatever light this might ultimately cast on *human* mentality. Admittedly, we would need to do a good deal more work to justify any step from "machine" mentality to human mentality: we would, for example, have to appeal to some *convergence* principle, to suggest that similar

[6]This is the distinction (insofar as there really is one) between H_t and the stronger doctrine of (unqualified) *Computationalism* (H_c): H_c claims that a computer which passes the Turing Test really does exhibit genuine mentality—that mentality just *is* some particular kind of computational activity.

reductions might be expected to apply to any systems, including human beings, exhibiting the relevant behaviour (Harnad 1989). However, since the level of machine "mentality" which has been achieved to date is extremely limited (falling far short of Turing Test performance) it seems a little premature to worry unduly about the ultimate scope of any such putative "machine psychology" at this stage.

Now, computational *universality* (e.g. Lewis & Papadimitriou 1981) guarantees that if a realisation of *any* abstract universal computer can pass the Turing Test (exhibit intelligence) when suitably programmed then, in principle, some (sufficiently "large" and/or "fast") realisation of *every* abstract universal computer can. Which is to say that the underlying programming formalism (which defines the abstract computer), once it is universal, does not *constrain* the intelligence of the machine. However, it *will* radically affect the *reduction* of mental to computational entities; this reduction will, at best, be unique only relative to a particular programming formalism.

Furthermore, we must at least recognise the possibility that the reduction may therefore be much *simpler* relative to some formalisms compared to others; and that which formalism is most illuminating may even be *different* depending on the particular aspects of mentality under consideration at any given time.

My point here is to distinguish between the *existential* and the *pragmatic* aspects of the Artificial Intelligence research programme. From an existential point of view, all (abstract) universal computers are equally powerful; if one can be made intelligent, they all can. But, from the pragmatic point of view, in terms of understanding or explaining intelligence (reducing the mental to the computational), there may be very substantial differences between universal computers, and, indeed, the most pragmatically useful computer may be different in different contexts.

To put it another way, consider the case of a reasonably good computer chess player—Dennett's (1978b, *passim)* prototypical example of a machine to which we can effectively adopt the intentional stance. One possible explanatory schema would be to attempt a *direct* reduction of the intentional attributions to characteristics of the program *as expressed in the native instruction set of its processor.* At best this will be unintelligible; at worst it will be hopelessly impractical. In-

stead, any effective explanation would certainly make use of a *hierarchy* of "levels" which *progressively* reduces (or explains) the intentional attributes. It is quite plausible that explanations at some of the different levels, or even within certain levels, may be effectively expressed in different formalisms (i.e. in terms of different *virtual machines*). It would be quite sterile to then argue about which of these formalisms is the "correct" one.

Compare also Dennett's (1978a) own discussion of a similar mode of explanation, which he describes in terms of progressively *discharging homunculi*, and Dawkins' (1986, p. 13) notion of *hierarchical reductionism*. While Dennett's discussion is most naturally applied to the decomposition of a program within a single programming formalism, similar principles would apply to transitions between different formalisms. More generally, while I have talked loosely about distinct programming "formalisms" and "virtual machines", the transitions need not always be clear cut or precise. There are hierarchies and hierarchies (compare also, the "Strange Loops" and "Tangled Hierarchies" of Hofstadter 1979).[7]

Converse arguments apply, of course, to the *synthesis* of intentional systems. While, "in principle", any abstract universal computer is as powerful as any other, in practice (i.e. in terms of the ease of designing the required functionality) some may be better than others, and a variety may be much better than any single one. The latter point still holds *even* if, ultimately, a system is *implemented* by simulating the required diverse computers on a single host. Note carefully that the distinction I am drawing here has nothing to do with the relative speed or performance of distinct, physical, computers: it is concerned purely with differences between programming formalisms, or abstract computers.

I am arguing here against a tendency to adopt an extremely over-simplified view of the computationalist thesis. Given Turing's (1950) results on universal computation, and the general notion of computationalism, there seems to be an almost overwhelming temptation to interpret H_c as positing the existence of some *specific* and *unique* (virtual) machine, or "language of thought", which, implicitly,

[7]Note incidentally that, even when one has full access to the design of an artefact such as a chess computer—i.e. when one can, in principle at least, adopt Dennett's "design" stance—the reduction of the intentional to the computational may still be a very difficult problem. For example, consider Dennett's comment on "innocently" emergent phenomena (Dennett 1977a, p. 107), or Hofstadter's discussion of "epiphenomena" (Hofstadter 1979, Chapter X).

is realised in (all) human brains, and which is a sufficient formalism for the direct explanation (reduction) of all cognitive phenomena. That is, that the architecture of cognition consists of a single significant virtual machine level, and even that this machine is of some particular class (such as, say, a LISP-like machine).

This kind of view can easily lead to essentially sterile argumentation about the "absolute" claims of particular abstract machines—say procedural versus declarative programming, or serial versus parallel programming, or "passive" versus "active" symbols.

More recently, this kind of argument has become an element of the on-going debate between "classical AI" and "connectionism". Consider, for example, the review of this debate presented by Fodor & Pylyshyn (1988). Their conclusion is that a connectionist "architecture", insofar as such a thing is well defined, is not a viable candidate as *the* "architecture of cognition". Now their arguments based on "combinatorial syntax and semantics" seem to me conclusive in this regard. In other words, contrary to some of the connectionist rhetoric, connectionism is not (or, at least, not necessarily) an *alternative* to classical ideas on cognitive *architecture*, but is, rather, *complementary* to it, particularly insofar as it may offer insight into ways to effectively *implement* certain aspects of classical architecture.[8] But even here there is a risk that Fodor and Pylyshyn could be (mis-?)interpreted as proposing that there does, in fact, exist *some* unique (though non-connectionist) programming formalism (abstract universal computer) which is *the* "architecture of cognition". Such an interpretation (which, I stress, may not be intended by Fodor and Pylyshyn) would be almost as bad as the position they attack: the proposal of connectionist networks as *the* "architecture of cognition".

This whole discussion is fraught with difficulty, and the possibility of misinterpretation. Thus, consider, for example, the *Physical Symbol System* hypothesis of Newell & Simon (which I will denote H_{pss}):

> A physical symbol system has the necessary and sufficient means for general intelligent action.
>
> Newell & Simon (1976, p. 41)

[8]Compare also Boden's slightly earlier review of this debate, where she considered, and ultimately rejected, the idea that connectionism might represent a Kuhnian "paradigm shift" relative to "Good Old Fashioned AI" (Boden 1988, Chapter 8).

Now Newell & Simon explicitly stipulate that a key element in their formulation of H_{pss} was the invention of LISP by John McCarthy, which became the prototypical example of a "symbol system", and the demonstration that such a system was "equivalent to the other universal schemes of computation". But if a *symbol system* is simply a particular *class* of (abstract) universal computer, and a *physical* symbol system is simply a realisation of a member of this class, then exactly how does H_{pss} go beyond the general computationalist hypothesis, H_c? Alternatively, if we accept the literal interpretation that H_{pss} posits that *only* a particular class of universal computers can exhibit general intelligence (i.e. the claim that a member of this class is *necessary*) then the hypothesis is simply false: by the definition of computational universality, as already discussed, if any (abstract) universal computer can exhibit intelligence then they all can.

Now, as a matter of fact, I believe that what Newell & Simon mean to claim by H_{pss} is (at least) that any intelligence system must have a "virtual machine" level which is an implementation of a symbol system (of a more or less LISP-like sort). This would a perfectly good qualification of H_c, i.e. a perfectly good additional hypothesis about the nature of "cognitive architecture", though it would need considerable clarification. But there again, Newell & Simon's claim may, in fact, be even stronger than this: it is simply very difficult to establish, unambiguously, what exactly they intend.

As evidence that this confusion is not merely an individual failing on my own part, consider Hofstadter's attempt at a critical evaluation of Newell & Simon's position (Hofstadter 1983); Newell dissented from this sharply in his accompanying commentary (Newell 1983), stating *inter alia* that Hofstadter was "mistaken, absolutely and unequivocally" (p. 293) in at least some of his interpretation; but Hofstadter has since repeated and reinforced his criticism, albeit with some clarification (Hofstadter 1985, Chapter 26, Post Scriptum).

There the direct argument currently rests, to the best of my knowledge. However, indirect reverberations continue. Thus, Fodor & Pylyshyn (1988, p. 59) are dismissive (verging on the sarcastic) in relation to the following particular passage from Hofstadter's original paper:

> The brain itself does not manipulate symbols; the brain is the medium in which the symbols are floating and in which they trigger each other. There is no central manipulator, no central program. There is simply a

> vast collection of "teams"—patterns of neural firings that, like teams of ants, trigger other patterns of neural firings. The symbols are not "down there" at the level of the individual firings; they are "up here" where we do our verbalization. We feel those symbols churning within ourselves in somewhat the same way we feel our stomach churning.
>
> Hofstadter (1983, p. 279)

Yet: precisely the same passage has been quoted, apparently *favourably,* by Boden (1988, p. 247).

As another example of ambivalence about the notion of an architecture of cognition, consider Fodor's book *The Language of Thought* (Fodor 1976). Reflecting the definite article used in the title, the book is dominated first by the attempt to establish that "the" language of thought exists, and then by the examination of what some of "its" properties must be.

Fodor starts off his argument with the statement that "representation presupposes a medium of representation, and there is no symbolisation without symbols ... In particular, there is no representation without an (*sic*) internal language" (p. 55). This could be interpreted simply as a variation on H_c, and, as such, is hardly objectionable; it is equivalent to the claim that a computer (generally) has a single *native* instruction set (language), in which every aspect of its behaviour can (in principle) be explicated. But: Fodor also claims that "a little prodding will show that the representational system ... must share a number of the characteristic features of real languages" (p. 31), and later he speculates specifically that "the language of thought may be very like a natural language ... It may be that the resources of the inner code are rather directly represented in the resources of the codes we use for communication" (p. 156). This kind of discussion strongly suggests that Fodor has in mind a single, unitary, language, distinct from any natural language, but not all *that* different, which is sufficient for the more or less direct reduction of mental phenomena. Now I emphasise that this simplistic view is only *suggested* by Fodor's treatment. Nowhere does he explicitly state anything to this effect; and there are at least some occasions when he appears to explicitly *reject* any such interpretation, such as the following:

> It is probably a mistake to speak of *the* system of internal representations that the organism has available for the analysis of environmental events or behavioral options. Rather, in the general case, organisms have access to a wide variety of types and levels of representation, and which one—or ones—they assign in the course of a given computation is determined by a variety

> of variables, including factors of motivation and attention and the general character of the organism's appreciation of the demand characteristics of its task.
>
> Fodor (1976, p. 157)

> ... we can see that 'the' representation that gets assigned to an utterance in a speech exchange must be a very heterogenous sort of an object. It is, in effect, the logical sum of representations drawn from a number of different sublanguages of the internal language. *It is an empirical question what, if anything, these sublanguages have in common ...*
>
> Fodor (1976, p. 159, emphasis added)

Now I agree wholeheartedly with this position; but I confess that I find it difficult to interpret the rest of the book in this light. Given such a view, it would seem that a sensible first step would be to emphasise *distinctions* between different systems of representation, rather than doing as Fodor does—which is to talk of a single system of representation (*the* language of thought) which encompasses everything.

In conclusion, I suggest that the application of the notion of universal computation in Artificial Intelligence could usefully be reversed from its usual formulation. It is usual to think of computational universality as indicating that arbitrary intentional phenomena can be "explained" in terms of a *single* mechanism—implicitly, that there is some theoretical gain ("parsimony"?) in doing so. By contrast, my view is that computational universality legitimises explanations which invoke arbitrarily complex combinations of different (computational) mechanisms, because we are guaranteed that, provided these are all effectively defined, they can all ultimately be "reduced" to a single mechanism (should there be any benefit in doing so).

That is, in considering the architecture of cognition, we need not conceive of that architecture as fundamentally identified with a particular *computer* (i.e. a particular "system" of representation, or homogenous "language of thought"); rather we may think in terms of an heterogenous *network* of (abstract) machines (homunculi) specialised for different tasks, which, in aggregation, might yield something approaching human intelligence.

This is not, of course, an original idea. For example, it is very similar to Minsky's *Society of Mind* (Minsky 1986), or Hofstadter's "soup cognition" (Hof-

stadter 1983). Similarly, Boden has recently emphasised the need to take a pluralistic view of computationalism, and to avoid over-simplification (Boden 1988, p. 232). Dennett has also ventured to provide a schematic design of computer "consciousness" which exemplifies these ideas (Dennett 1978c). Perhaps then I have laboured the issue unduly; and yet, I think it is clear from the literature I have reviewed above that this central point—the potential for a programmed universal computer to *simultaneously* admit descriptions both of trivial simplicity, and almost inconceivable complexity—has *not* been consistently recognised, or clearly enunciated.

The point is, in any case, crucial for my purposes here. Specifically, the work to be presented in subsequent chapters may seem, by comparison with AI research generally, to be of a relatively primitive sort; but my claim is that the problem of building, and understanding, an artificial intelligence is of unknown, but certainly vast, proportions, and that it is only with an appreciation of this that the need for the kind of fundamental research described here can be properly understood. To underline this, I close this discussion of cognitive architecture with a final quotation from Fodor:

> On the one hand, internal representations are labile and the effectiveness with which they are deployed may, in given cases, significantly determine the efficiency of mental processing. On the other hand, we know of no general constraints on how information flows in the course of the computations which determine such deployments: To say that we are dealing with a feedback system is simply to admit that factors other than the properties of the input may affect the representation that the input receives. In particular, what internal representations get assigned is sensitive to the cognitive state—for all we know, to the *whole* cognitive state—of the stimulated organism. Perhaps there are bounds to the options that organisms enjoy in this respect, but if there are no one now knows where to set them. Psychology is very hard.
>
> Fodor (1976, p. 166)

3.5 On Computational Semantics

The essence of the Turing Test, as already discussed, is to judge whether a system *means* what it says, or *knows* what is being talked about—in a comparable sense to the way we use these terms for human beings. This is considered to be diagnostic of intelligence, subsuming other aspects, such as consciousness, creativity,

imagination, etc.—presumably because these latter things would, implicitly, be tested anyway: they are so central to our granting that the system understands our conversation at all that we would surely, *inter alia*, insist on testing whether it understands these particular concepts themselves.

So, the key question which arises in relation to a programmed computer which passes the Test, is: what is the relationship between its knowledge and the formal entities (tokens) making up its program?[9] Or, equally: which of these tokens mean anything at all, and, of those which do mean something, just what do they mean? Or, in the terms of the discussion of the previous section, we are asking: how are the mental entities *meaning* or *knowledge* to be reduced to the tokens constituting the computer's program?

Note again that any answer to these questions must be contingent on the definition of the computer itself: on its being the particular universal computer postulated by the program. To this extent, the reduction of meaning or knowledge to aspects of the program would not represent the complete reduction to physical terms: however, I take it that the reduction of the computer itself to physical terms will not be problematic.

We should feel much happier in attempting to answer these questions if we already had available to us a detailed specification of a programmed computer which *can* pass the Turing Test. Unfortunately, of course, we do not. Indeed, the fact is that the very design or construction of such a system presupposes, to some extent, that we already *know* what the answer to these questions is. More precisely, we must hypothesise answers to these questions as a prerequisite to building a computer system which could pass the Turing Test.

So, we need a theory of meaning or semantics, applicable to computational systems in general.

On the face of it, there is no shortage of competing theories for this purpose. I shall certainly not attempt a comprehensive review here. However, I shall suggest

[9] I use "program" without prejudice to the programming formalism, and eschewing any distinction between "instructions" and "data"; to put it another way, to the extent that any practical digital computer has finite storage, "program" can conveniently be interpreted as synonymous with the *state* (of this finite state machine). I use "token" to denote any arbitrary component of a program. It need not, and generally does not, imply any kind of "atomic" or "primitive" component of a program: indeed the entire program will be considered to constitute one particular token.

that, despite differences in vocabulary and emphasis, there is a common core to several of the theories of meaning which are in practical use within AI. I shall try to identify, and make more explicit, this common core, and to then use it as a foundation for subsequent developments.

I must consider first the view that there cannot be a "computational semantics" at all—that computers (and/or their programs) simply are not the kinds of things to which, in whole or in part, meaning can be ascribed. This position has been put forward by a number of writers. The basic idea is that a computer program is no more and no less than a formal, or syntactic object, and, as such, cannot intrinsically *refer* to anything. It may (or may not) be possible to systematically *interpret* it as referring to something, but this meaning, or understanding, is entirely in the mind of the person doing the interpretation—it is not, and cannot be, a property of the program "itself".

There is, of course, a grain of truth in this view. Thus, we may compare a computer program to a book; the book has meaning only insofar as people read and understand it. The book does not understand itself.[10] Or we may identify a program with its *information* content—in the sense of the formal theory of information; such information content would be independent of the meaning of the program (if any). Indeed, one might underline this by arguing that any particular *arbitrary* sequence of characters, of the same length as the program, would literally have the same information content (relative to an implicit, equiprobable, ensemble of all such sequences).

I said that there was a grain of truth in this view, and it is this: a program cannot be given meaning simply by wishing it so. In particular, one cannot cause a program token to mean something merely by giving it a particular label—i.e. the use of what McDermott (1976) calls "wishful mnemonics". This also applies of course *mutatis mutandis* to the tokens output by a program—as artfully, if accidentally, demonstrated by Weizenbaum with his infamous `ELIZA` program(s) (Weizenbaum 1984; see also Dreyfus & Dreyfus 1986, Chapter 3).

[10] Popper diverges somewhat from this common-sense view, with his idea of the World of *objective knowledge* (World 3). However, I do not think this is critical for the issue at hand here: while Popper would challenge the view that the only kind of knowledge is (subjective) knowledge of organisms, especially people, he certainly would not claim that a book literally understands itself.

Equally of course, the fact that a token of a program has a particular label does not necessarily mean that it is *devoid* of meaning: it merely says that *whatever* (if any) meaning it has is not *by virtue* of the label.

The *formalist* idea—that a computer program, being a purely formal object, cannot really mean anything—has been reviewed in detail, and rejected, by Boden (1988, Chapter 8). I consider that her analysis is correct, and I do not propose to repeat her detailed arguments here. The essential point is that, although a computer program *per se* may be viewed as a purely formal object, one cannot say the same for a programmed *computer*. The computer does actually *do* something (as a consequence of its program—but also depending on its inputs). This is just to repeat the point made earlier that, when one ascribes mentality to a program, this is always shorthand for referring to the system consisting of a (compatible) computer which is actually running the program. The shorthand is reasonable because we assume that the computer itself can be "easily" reduced to a lower (physical) level, should one wish to do so: the really complicated or interesting phenomena are evidently related to the particular program, rather than the computer. But while the shorthand is reasonable, it is open to misinterpretation: specifically as being a claim that a program has *intrinsic* meaning, of and in itself, *independently of any particular computer*. This latter idea is certainly mistaken, and, insofar as this is the object of the formalist's attack, the attack is justified. But the point remains that this cannot be turned into a general attack on the idea that programmed computers can have "genuine" semantics.

To return now to this main theme: following Boden, I discount the suggestion that the formal tokens making up a computer program (embedded in a suitable computer) *cannot* have "intrinsic" meaning. So far so good, but I have not yet made any positive suggestion as to what it could or should mean (!) to say that a program token *does* mean something in this intrinsic sense.

To progress this, let me first consider a terminological point. I have generally used the word *token* to refer to the constituents of programs (regardless of what the programming formalism might be), rather than the word *symbol* (except where I was citing other authors who have, or at least *seemed* to have, adopted the latter usage). My reason is, of course, that *token* doesn't prejudice the issue of semantic content, whereas once we describe something as a *symbol* we are im-

plying that it has meaning, that it refers to something, that it serves to *symbolise*, as least with respect to *some* interpreter.

Now the conventional notion of a symbol places some emphasis on its *arbitrary* or *conventional* character: a symbol is viewed as a vehicle of communication (or, perhaps, of memory), and, as long as the communicating parties are agreed as to its meaning (even if only approximately), then the exact nature of the symbol is largely irrelevant. Of course, it is admitted that a symbol *may*, in some sense, "resemble" the thing symbolised, but this is not considered necessary or criterial for its being a symbol. This is all true and valid, but I think it may be misleading. It puts the emphasis entirely in the wrong place. While it is true that the symbol, viewed in *isolation* can be entirely unrelated to its referent (which is merely to reiterate yet again that the symbol, in isolation, *is* meaningless), the symbol, viewed in *context*, *must* be related to the referent. That is, a symbol *is* a symbol because it *is* related (somehow) to the referent.

Informally, my general claim is that what makes a token a symbol is that it interacts with some system(s) in a manner which is related, in some more or less definite (though perhaps very complex) way, to the manner in which the *referent* interacts with the same system(s). More concisely, a token is a symbol when it is used by some system to *model* something.

There is, of course, nothing novel or original in this; it is merely an attempt to spell out the implication of calling something a symbol—namely that a token is only ever a symbol *in relation to some system*, and that, further, its referent must bear, in some (identifiable) sense, a similar relationship to the system. Dennett puts it thus:

> ... nothing is intrinsically a representation of anything; something is a representation only *for* or *to* someone; any representation or system of representations thus requires at least one user or interpreter of the representation who is external to it.
>
> Dennett (1978b, Chapter 7, p. 122)

This is worth spelling out in detail because, in dealing with computer programs, it is all too easy to confound two senses in which a token can symbolise: it can symbolise something to the *programmer* and it can symbolise something to the (rest of) the *computer*. The former sense of symbolisation is, of course, the

basis for the formalist argument that program tokens have no "intrinsic" meaning. The argument is wrong, but it is a very natural misunderstanding. Consider, as a more or less random example, the following quotation from Boden:

> No one with any sense would embody list-structures in a computer without providing it also with a *list-processing* facility, nor give it frames without a *slot-filling* mechanism, logical formulae without *rules of inference*, or English sentences without *parsing procedures.*
>
> Boden (1988, p. 249, original emphasis)

While Boden is here arguing for exactly the same point as I am attempting to make, it seems to me that she can easily be misunderstood. My point is that list-structures (say), in the *absence* of a list-processing facility, are only list-structures by courtesy; that is they are only list structures *relative to a human interpreter* (the programmer or otherwise). But, in the *presence* of a list-processing facility, then they become list-structures *relative to that facility.* To avoid confusion one should ideally always consistently refer to the meanings of the tokens only relative *either* to the computer *or* to a human interpreter; but, in any case, one should not switch between these two viewpoints without warning or comment, as Boden does here. The situation is not, as Boden puts it, that no one "with any sense" would embody list-structures without a corresponding list-processing facility; rather, in the sense in which Boden evidently means (i.e. relative to the computer) no one, sensible or otherwise, *could do it*—it is absolutely not a matter of choice. The idea of "embodying" list-structures in a computer without a corresponding list-processing facility is literally a contradiction in terms. The tokens in question, whatever they might symbolise to the programmer, cannot be said to symbolise (or to be) list-structures *relative to the computer* except in the case that *it* treats them so: i.e. that it *has* corresponding list-processing facilities.

Boden's version seems to be intelligible only if we interpret her to mean that, in the absence of a list-processing facility, the meaning of the relevant tokens (i.e. that they are *inter alia* list structures) should be interpreted relative to a human interpreter; *but* that, in the presence of a list-processing facility we can (and should?) change our viewpoint and interpret their meaning relative to the computer instead. But, if this *is* the interpretation Boden intends, it is poorly expressed.

Again, let me stress that this example was picked virtually at random, and I do not intend any particular or individual criticism of Boden. The problem is intrinsic to the nature of software *engineering*, and is very difficult to avoid. McDermott's idea of the "wishful mnemonic", already discussed, is obviously closely related to this.

I doubt that this point can be overemphasised. In any discussion of the "meaning" of the tokens associated with a computer program it seems that the *only* way to stay properly honest is to consistently ask "meaning relative to whom?"; and to then absolutely restrict our attributions of meaning to those which are valid or demonstrable *relative to the (programmed) computer* (or some subsystem thereof—such as some (other) specified tokens of the program). That is, we must constantly resist the temptation to, as it were, *anthropomorphize* meaning into a token.

I may add that I consider that this discipline, properly applied, seems to yield a unitary basis for a computational semantics, which resolves, or at least neutralises, the so-called *dual-calculus* view of computer programming. This view posits a sharp division of a program into active "instructions" and passive "data". Such a view leads to entirely misguided attempts to more or less exclusively or independently attribute meaning to specific aspects of instructions or data *separately.* But in fact, this distinction, while of considerable practical value in conventional software *engineering* (i.e. the development of software to exhibit particular, *effectively specified,* behaviours), lacks any intrinsic theoretical foundation.[11] That is, whether a human observer chooses to describe a particular token as "instruction" or "data", or as being "active" or "passive", or "declarative" or "procedural", is strictly irrelevant to its meaning (if any) *to the computer.* The latter meaning (which is the only meaning of interest in the AI context) can *only* be established

[11] As discussed by Hodges, in his biography of Turing, this point was already implicit in Turing's invention of the Universal Turing Machine, though it was not explicitly recognised by Turing until he set about designing his first practical digital computer (Hodges 1983, Chapter 6, pp. 324–327). It was documented in Turing's report on the *Automatic Computing Engine,* or ACE (c. 1945). As Hodges put it: "It was ... a small step to regard instructions ... as grist to the ACE's own mill. Indeed, since so much of his [Turing's] war work had depended upon indicator systems, in which instructions were deliberately disguised as data, there was no step for him to make at all. He saw as obvious what to others was a jump into confusion and illegality." Hodges also notes that this insight was not explicitly pointed out "on the American side" until 1947.

by reference to the objective interactions or effects of the token on the rest of the system (i.e. the programmed computer) in which it is embedded.[12]

In any case, to return to the question of *what* a token means (as opposed to the question of *to whom?*), we may distinguish two cases. Trivially, we may identify the meaning of the token with its *direct* effect on its interpreter. Thus we might say that a particular token "means" precisely that its interpreter should do whatever it is it does in response to that token. This notion of meaning is possible (if not very helpful) for very simple, determinate, interactions. But in more complicated cases we want to identify the meaning of the token, to its interpreter, with some aspect of the interpreter's *environment.* This is the normal usage of *symbol* or *reference*: the token *refers* (in the sense of being taken by the interpreter to refer) to some (other) thing in the interpreter's environment. And this brings us back to the notion of the symbolic token as a *model.*

This general notion of model based semantics is well established; but there is some room for debate, if not disagreement, as to what we should admit as a "model". For my purposes it is not essential to tie this down too precisely. Instead, I shall review and compare some selected ways in which it has previously been applied.

The most comprehensive, and mathematically rigorous, review of the modelling relationship of which I am aware is that of the mathematical biologist Robert Rosen (1985a). I shall therefore base my presentation of a computational semantics on Rosen's concept of an *anticipatory system*:

> An anticipatory system S_2 is one which contains a model of a system S_1 with which it interacts. This model is a predictive model; its *present* states provide information about *future* states of S_1. Further, the present state of the model causes a change of state in other subsystems of S_2; these subsystems are (a) involved in the interaction of S_2 with S_1, and (b) they do not affect (i.e. are unlinked to) the model of S_1. In general we can regard the change of state in S_2 arising from the model as an *adaptation*, or pre-adaptation, of S_2 relative to its interaction with S_1.
>
> Rosen (1985a, p. 344, original emphasis)

Where a relationship of this sort exists, I shall say that the subsystem of S_2 which is the predictive model of S_1 *means* or *refers* to S_1; and that, in this

[12]This dual calculus issue has again been thoroughly reviewed by Boden (1988, Chapter 8, pp. 248–250), though not from quite the perspective suggested here.

sense, S_2 *understands* or has *knowledge* of, S_1. I take it that, in the case of direct interest here, i.e. where S_2 is a *computational* system, then the relevant subsystem of S_2 (the predictive model of S_1) will be identifiable with a particular token of S_2's program (although this need not be the *only* function of this token, or all its components). I repeat that such "tokens" will, in general, be composite, dynamic, objects. This token then *means* or *symbolises* S_1 (to S_2); it is, genuinely or intrinsically, a *symbol.*

I do not insist that these are absolutely *necessary* conditions for meaning or grounding; but it seems to me that they may be *sufficient.*

Consider now Newell & Simon's description of what it is for a formal token to refer to, or mean something:

> A symbol structure *designates* (equivalently, *references* or *points to*) an object if there exist information processes that admit the symbol structure as input and either:
>
> (a) affect the object; or
>
> (b) produce, as output, symbol structures that depend on the object.
>
> Newell & Simon (1972, p. 21)

Condition (a) here is comparable to the requirement that S_2 (which encompasses at least the symbol structure in question and the specified information processes) must be capable (potentially, at least) of *interacting* with S_1; condition (b) is comparable to the requirement that S_2 does, in fact, contain a model of S_1. Now Newell & Simon state only these two conditions, and state them as alternatives; whereas I require both conditions, and additionally stipulate that the putatively symbolic token ("information process") must be *predictive.* Thus it is seen that my conditions for meaning or grounding are compatible with, but rather more severe than, those suggested by Newell and Simon.

I now turn to an early discussion by Dennett, in his *Content and Consciousness* (first published in 1969), where he considers the problem of the ascription of content to (physical) states and events in brains (see Dennett 1986, Chapter IV). While Dennett's concern here is mainly with human psychology, or "real" mentality, rather than with the putative mentality of a suitably programmed computer, his concepts are presented as being applicable to *any* "intentional" system, and

should thus carry over more or less directly. Dennett's treatment is quite technical and involved, but is roughly summarised by the following extract:

> The content, if any, of a neural state, event or structure depends on two factors: its normal source in stimulation, and whatever *appropriate* further efferent effects it has; and to determine these factors one must make an assessment that goes beyond an extensional description of stimulation and response locomotion. The point of the first factor in content ascription, dependence on stimulus conditions, is this: unless an event is somehow related to external conditions and their effects on the sense organs, there will be no grounds for giving it any particular *reference* to objects in the world. At low enough levels of afferent activity the question of reference is answered easily enough: an event refers to (or reports on) those stimulus conditions that cause it to occur. Thus the investigators working on fibres in the optic nerves of frogs and cats are able to report that particular neurons serve to report convexivity, moving edges, or small, dark, moving objects because these neurons fire normally only if there is such a pattern on the retina. However mediated the link between receptor organ and higher events becomes, this link cannot be broken entirely, or reference is lost.
>
> The point about the link with efferent activity and eventually with behaviour is this: what an event or state 'means to' an organism also depends on what it *does* with the event or state ... Where events and states appear inappropriately linked one cannot assign content at all, and so it is possible that a great many events and states have no content, regardless of the eventual effect they have on the later development of the brain and behaviour.
>
> Dennett (1986, Chapter IV, pp. 76–77)

I suggest that this position of Dennett's is closely related to the position I have already described in relation to anticipatory systems, although Dennett provides some useful complementary insights also.

Taking first Dennett's discussion of the relationship between neural entities (or tokens, in a computational system) and "stimulus conditions", this mirrors my earlier requirement that the token be a *model* of the thing referred to. Dennett's version is somewhat more restrictive, and I would argue that it is unnecessarily so: for a subsystem might, conceivably, operate successfully as a model *without* any ongoing *linkage* to the thing modelled (i.e. without any ongoing linkage to "stimulus conditions"). However, I would grant that some such linkage to stimulus conditions may be a necessary factor in the *original* development or establishment of any modelling relationship; my point is simply that, once the modelling relationship is established, it *may* successfully persist, even if the original linkage

to stimulus conditions is broken. It may also be that, given the context of Dennett's discussion, he was primarily making a *pragmatic* rather than a *theoretical* point: i.e. that only modelling relationships which (still) have extant linkages to stimulus conditions would be retrospectively identifiable in practice. Even this would be a debatable claim however. A separate point arises from my claim that the model should be *predictive*; Dennett certainly makes no explicit statement of this sort. However, his requirement that the ultimate efferent effects be *appropriate* might, conceivably, be argued as amounting to the same thing.

Moving on to the efferent effects, Dennett's general requirement that there must *be* some such effects, and that they must be *appropriate* corresponds quite well in my formulation to Rosen's stipulation that the predictive model embedded in the system S_2 must actually affect the interaction between S_2 and S_1. That is, the model must have some effect (S_2's behaviour must depend on the model), and a minimal constraint on the appropriateness of the effect is that it be concerned with the interaction with S_1. Crudely, if S_2 has a model of S_1, but only actually "uses" it in its dealings with some other (unrelated) system S_3, then we could hardly describe such usage as "appropriate". Again, Dennett's position may be somewhat more restrictive than is captured by the notion of anticipatory system: in particular, Dennett may well have in mind some more stringent tests of the appropriateness of behaviour (i.e. stronger than that the behaviour just be directed at the "right" target). I would accept that some such stronger tests might be useful in the case of biological systems, but I am not convinced that they are necessary in the general case.

In any case, it should be clear that the discrepancies, such as they are, between Dennett's view of the ascription of content and my discussion based on anticipatory systems, all rest on rather fine distinctions, and are not fundamental. In fact, I would suggest that the general notion of an anticipatory system satisfactorily captures Dennett's own idea of an intentional system, but in relatively more formal terms; that is, it seems to me that modelling a system as being anticipatory (relative to some environment in which it is embedded) is virtually synonymous with adopting the intentional stance toward it.

Next I shall consider Boden's (1988) review of the issue of computational semantics. She considers, in particular, Montague's (1974) *model-theory*:

> Model-theory deals with how meaning can be assigned to an uninterpreted formal system. Broadly, it says that such a system can be interpreted as being *about* a given domain if that domain can be systematically mapped onto it. If there are two such domains, the formal system in itself is about the one just as much as it is about the other.
>
> Boden (1988, p. 131)

It is clear that the general principle here is compatible with the view I have been describing in relation to anticipatory systems. However, it embodies only the condition that a modelling relationship must exist: it does not stipulate a *predictive* model, and, more importantly, does not require that the model must have effects on the interaction with the thing modelled. It is precisely this latter omission which introduces an unnecessary extra degree of ambiguity in the ascription of meaning. That is, if S_2 has a model which, in fact, models both S_1 and S_3, but which affects only the interaction between S_2 and S_1 (indeed, it may be that S_2 does not interact with S_3 at all), then I would argue that the model should only be said to be about S_1, and is definitely not about S_3—whereas Montague's theory would appear to admit both ascriptions equally. This is not to suggest that the anticipatory model eliminates all ambiguity of meaning: there certainly may be cases in which a model is about (affects the interaction with) more than one referent. Rather, I am saying that the conditions I propose for admitting such ambiguity are significantly more restrictive than those accepted by Montague.

It seems to me that, although this discrepancy is not too serious in itself, it may actually be symptomatic of a more fundamental difference. The fact is that, despite the suggestive overlap in vocabulary between Montague and myself, and the fact that Boden introduces Montague's work in the context of computational psychology, Montague is actually dealing with a different problem from that with which I am concerned.

My reading of Boden is that Montague is concerned with whether, in general, *isolated* formal systems can be said to mean anything. This is the problem to which his model-theory offers an answer. If I am correct in this interpretation, then I should argue that the answer is incomplete, if not actually mistaken.

This comes back to the question of *meaning to whom?* If we are really dealing with a completely isolated formal system, then I assert that it really is meaningless, regardless of what mappings may arguably exist between it and arbitrary domains. If, on the other hand, we do not really mean that the formal system is isolated, but are simply saying that it *could (potentially) be* interpreted as mapping onto certain domains (although it is not, as it were, currently or actively being so interpreted), then the theory becomes coherent but incomplete: it could only be completed by stipulating a set of (potential) interpreters.

The point is that, in the former case, if Montague is dealing with truly isolated formal systems, then his model-theory (whether right or wrong) is irrelevant—its application in my context would involve a throwback to seeing a computer program as being *purely* a formal object, a perspective I have already rejected. Conversely, in the latter case, if Montague is dealing with meaning relative to *sets* of interpreters, then, for my purposes, the theory is too general—for there is only one interpreter with which I am immediately concerned, and that is the specific system (the programmed computer) in which the formal system is embedded. With respect to *that* interpreter, I argue that the more specific (and restrictive) theory of meaning based on anticipatory systems, which specifically takes account of this particular interpreter, subsumes whatever applicability Montague's theory might otherwise have had.

Boden does not pursue this issue in depth, but it seems that her ultimate conclusions are, at least, not incompatible with my analysis. Thus, Boden ultimately merges this discussion with a more general discussion of whether a computational semantics is possible at all. There she comments that "some writers [argue] that computer programs have an intrinsic causal-semantic aspect (*to be distinguished from any abstract isomorphism there may be between programmed formalisms and actual or possible worlds*)" (Boden 1988, p. 238, emphasis added). This seems to me to be essentially the same point I have tried to make above in relation to the applicability of Montague's theory. Boden concludes:

> In a causal semantics, the meaning of a symbol (whether simple or complex) is to be sought by reference to its causal links with other phenomena. The central questions are "What causes the symbol to be built and/or activated?" and "What happens as a result of it?"
>
> Boden (1988, p. 250)

At this point, Boden's review is completed, but it should be clear that it has more or less met up with Dennett's discussion, already dealt with above. The relationship with the theory based on anticipatory systems is therefore similar, and I shall not detail it again.

I should now like to consider Harnad (1990), who introduces what he calls the *symbol grounding problem.* This is closely related to, but not identical with, my problem of the conditions under which computational tokens may be said to have meaning or reference. Harnad's problem is not identical with mine because Harnad *accepts* Searle's Chinese Room argument, and concludes from this that there are certain relatively severe constraints on symbol "grounding". I, on the other hand, reject Searle's argument (see Chapter 2 above, section 2.4.1), and therefore also reject the inferences made by Harnad from that argument. Notwithstanding this difference between us, I think it worthwhile to review Harnad's actual grounding proposals.

In this discussion, it is important to note that Harnad uses *symbol* in the special sense of a token embedded in a system of tokens which admits of "semantic interpretability" or *systematic* interpretation. That is, the meaning of any composite token (symbol) can be effectively established in terms of the meanings of its atomic components. In Harnad's terms then, the problem is that while both people and "symbol systems" can analyse the meaning of a composite token in terms of the meanings of its components, and may (via a "dictionary" or otherwise) be able to further analyse this in terms of the meanings of some (smaller) set of *primitive* atomic tokens, these primitive tokens are then, in themselves, meaningful for people but not for symbol systems. For these tokens their meaning is not (by definition) a consequence of some "definition" in terms of other tokens, so what can it be?

Harnad's outline answer is that such tokens will be meaningful or well grounded if they are derived from, or causally related to, "non-symbolic" representations of their referents, specifically in the form of what he terms *iconic* and *categorical* representations.

Iconic representations are "internal analog transforms of the projections of distal objects on our sensory surfaces" (Harnad 1990, p.342). It is quite difficult to tie this kind of idea down precisely—one immediate problem is specifying

how far into the nervous system we can still speak of a "sensory surface". But roughly speaking, Harnad means something like an *image* in the case of vision, and something analogous to this for other sensory modalities. So, an icon is something that more or less matches the sensory projection which an object would make (whenever present in the "immediate" environment) at some more or less well specified "level" or "locus" in the nervous system.

Iconic, or imagistic, representations are a well established notion in theories of mentality. I believe that it is now generally accepted that such representations (and processes constrained by them) certainly cannot account for all aspects of mentality, but it seems, equally, that they do play *some* important roles (however, see Dennett 1986, Chapter VII, and Dennett 1977b, for some critical discussion of the issue). And, of course, that is exactly the scope of Harnad's proposal—not that iconic representations are the sole vehicles of mentality, but that they do play at least one critical role, namely being an essential step in the grounding of (primitive) *symbols* (in Harnad's sense of that word).

A *categorical representation* of an object is then a derivative of an icon, which has been "selectively filtered to preserve only some of the features of the sensory projection: those that reliably distinguish members from nonmembers of a category" (Harnad 1990, p.342).

Harnad argues that iconic representations are necessary to allow *discrimination* of sensory inputs; that categorical representations are necessary to allow *identification* of objects (categorisation) in sensory input; and that both of these are necessary, though not sufficient, underlying processes affecting a "symbol" in order for that symbol to be well grounded. The further conditions that are sufficient for the symbol to be well grounded are that the complete system embodying the symbol must be able to "manipulate", "describe",[13] and "respond to descriptions" of the objects referred to. He does not go into detail of how this might be achieved, but appears to suggest that general symbol manipulation capabilities (universal computation ability operating upon the grounded symbols?)

[13]Harnad specifies that human beings can both "describe" and "produce descriptions of" objects; he clearly intends some distinction between these two, but I have been unable to understand what it is.

are the only *additional* facilities required, together with an assumption that "appropriate" processing is realised by these facilities (i.e. they have been suitably programmed?).

In terms of my own presentation of meaning in the context of anticipatory systems, these ideas of Harnad's fit in reasonably well. Iconic representations certainly are a kind of *model*, particularly suited to certain kinds of usage (namely discrimination); categorical representations are another kind of model, particularly suited to other kinds of usage (namely identification). To this extent, Harnad's proposals are compatible with mine, but are more detailed. However, Harnad appears to insist that these are *essential* components in the grounding of any symbol (whether directly, as in the case of primitive symbols, or indirectly for all others). I suggest that this condition is too strong: certainly, *some* modelling relationship is essential to grounding (to establish reference at all), but I don't see that certain particular forms of such relationship, such as those singled out by Harnad, are uniquely necessary.

More generally, I would argue that iconic and categorical representations are a very crude and limited form of model, and I suggest that they are, at best, the tip of the iceberg of general "symbol grounding".

To the extent that Harnad emphasises modelling relationships most closely related to processing of sensory input, his position is, perhaps, not dissimilar to that of Dennett. As we have seen, Dennett emphasises the need for linkage with sensory input as a basis for ascribing content. So, again, the point I wish to make is that while a close relationship to sensory input is one particularly plausible basis for establishing and/or recognising modelling relationships, it seems to me that it is not generally a *necessary* condition for the existence of such relationships.[14] I emphasise, of course, that I do not rule out the involvement of sensory input in modelling relationships; I simply stress that, in general, it may not be essential.

As to Harnad's remaining criteria for grounding (manipulation, description etc.), I suggest that these can be viewed as specific forms of my general require-

[14]Indeed, an undue reliance on sensory linkage might pose serious problems about the development of coherent mental activity despite very limited sensory ability (see Popper's remark regarding Helen Keller, quoted below). More speculatively, this is related to the persistence (at least for limited periods) of mental activity despite sensory deprivation (compare also the science fictional "brain in a vat" kind of question—e.g. Dennett 1976). However, these are extremely complex issues which I shall not attempt to discuss in detail here.

ment for the token (the predictive model) to affect the interaction with the object referred to.

As a final comment here on Harnad's discussion of symbol grounding, I note that he specifically cites Fodor (1976) as introducing semantic interpretability as criterial (or perhaps even sufficient?) for a "token" to be a "symbol". The subsequent thrust of Harnad's analysis is to reject such a purely "symbolist" view of symbol grounding, culminating in his introduction of iconic and categorical representations as alternatives to this view. Now, I have already noted, in the previous section, that Fodor could be misinterpreted in his book as proposing a unitary language of thought, at least on a superficial or cursory reading; and this seems to be precisely what Harnad has done. But, that it *is* a misinterpretation (or, at least, an oversimplification) should be clear from the fact that Fodor included, in the book, an extended and positive discussion of the use of *images* (icons?) and *discursive descriptions* of images (categorical representations?) (Fodor 1976, Chapter 4, pp. 174–194). This closely parallels Harnad's discussion, yet Harnad makes no reference to it.

To complete this sketchy review of computational semantics, I turn finally to Popper. As discussed in Chapter 2, Popper is no friend of physicalism, or, more especially, of computationalism. However, notwithstanding this, I wish to argue that Popper's *general* epistemology is compatible with the theory of meaning or reference being propounded here (contrary perhaps, to Popper's own wishes and beliefs), and can, indeed, serve to illuminate this theory further. This is a rather important point to establish, since I shall be drawing heavily on Popper's work in subsequent sections, particularly in relation to the *growth* of knowledge.

Popper's concepts of the Worlds 1, 2, and 3 have already been introduced in Chapter 2. The point argued there related to the reducibility or otherwise of World 3 to World 2 (and, in turn, of World 2 to World 1). Popper argues for their irreducibility, and in this sense, he envisages that there exists some form of "knowledge" which is *not* accessible to, or realisable by, computers. By contrast, I have claimed that Popper's argument is flawed, and thus it does not follow that what I call "computational semantics" is *necessarily* impoverished in some sense. However, I do not want to reopen that particular debate here; my immediate objective is more modest, and will be pursued separately. I wish to establish,

firstly, that there is *some* sense in which Popper admits that machines, such as computers, can realise or embody "knowledge"; and secondly, that *this* kind of knowledge is essentially equivalent to the various formulations of a computational semantics which I have already discussed above.

Popper himself has not, as far as I am aware, given *detailed* consideration to the question of whether, or how, his epistemology can be applied to machines in general, or computers in particular; that is, he has not dealt very explicitly with the application of "knowledge" or "knowing" to computers and/or their programs. However, we may glean a satisfactory insight into his views by examining a variety of his writings.

There is first a discussion (dating originally from the period c. 1951–1956) in which Popper introduced the idea of a "predicting" machine, and spoke loosely in terms of its being "endowed" with, or being an "embodiment" of, knowledge (Popper 1988, pp. 68–77); granted, within that same discussion, Popper explicitly cautioned that he should not be taken as subscribing to the doctrine that "men are machines"; but, to repeat, that is not the issue just here.

Popper has consistently stressed the continuity of the biological world—that his idea of *subjective knowledge* (at least) allows for some kind of continuum, linking all living things. For example, he notes that subjective knowledge "should better be called organismic knowledge, since it consists of the dispositions of organisms" (Popper 1970b, p. 73). Now he has not explicitly used this phrase (subjective knowledge) in relation to machines, but he has, in a discussion of biological evolution, which relates specifically to the development of subjective knowledge, actually used a machine, *in place of a living organism,* to illustrate his argument (Popper 1961, pp. 274–275). Granted, Popper there emphasises that he leaves open the question of whether "organisms are machines"; but it is clear that, to some extent at least, he allows that his notion of subjective knowledge may be applicable to machines.

An underlying issue here is the (apparent) distinction between *computers* and *robots.* Thus, while Popper is dismissive in general of the "intelligence" of computers *per se,* he seems less definite about robotic systems. Specifically, the hypothetical machine referred to above, which Popper actually describes, *inter alia,* as a "complicated organism", is, in fact, a robotic aircraft. Popper even

goes so far as to refer to this machine's " 'mind' " (the inner scare quotes are, however, Popper's own).

This notion—that there is some fundamental distinction between the capabilities of computers *per se* and (computers embodied in) robots—is not uncommon. As already mentioned in section 3.2.3, French (1990) has viewed android capability as essential even to the passing of the conventional, strictly linguistic, Turing Test. For somewhat different reasons, relating to Searle's Chinese Room argument, Harnad (1989) has proposed what he calls the *Total Turing Test* which explicitly calls for full android imitation of human abilities. Boden (1988, pp. 242–245) has also taken robotic embodiment (of a computer) as a decisive element in responding to the Chinese Room argument (though her response is somewhat different from that of Harnad). A variety of other workers have, on more general grounds, advocated some form of robotic embodiment as a more or less essential aspect of realising AI (e.g. Dennett 1978d; Brooks 1986; Beer 1990; Cliff 1990).

This is a difficult issue, which I do not propose to discuss in depth here. I shall, however, state a position. It seems to me that any given system (computer or otherwise) may, potentially, be linked or coupled with its environment in an indefinitely large number of distinct ways or modalities, perhaps even with a continuum of alternatives in between these modalities. I do not doubt that the exact nature of these linkages affects the potentialities of the system. But I hold that we have, as yet, very little if any theoretical understanding of these phenomena; and, in particular, we have not yet got any basis for making a *fundamental* distinction between systems having only, say, a purely "linguistic" (VDU or teletype) interface, and systems having more extensive "human-like" or "robotic" linkages.

Returning to Popper, I may say that the apparently suggestive implication of his choice of a robotic machine rather than something closer to an isolated computer is, in any case, largely nullified by his constant rejection of "sense data" as critical for, or (worse) constitutive of, knowledge. Thus, for example, we have the following analysis:

> According to psychological sensualism or empiricism, it is the sensory input of information on which our knowledge and perhaps even our intelligence depend. This theory is in my opinion refuted by a case like that of Helen Keller whose sensory input of information—she was blind and deaf—was

> certainly far below normal, but whose intellectual powers developed marvellously from the moment she was offered the opportunity of acquiring a symbolic language.
>
> Popper & Eccles (1977, Chapter P4, p. 124)

Incidentally, Turing made much the same point at an earlier date, also citing the example of Helen Keller, and specifically applied this in the context of AI (Turing 1950, p. 456).

Thus far, I have simply argued that Popper should not be read as claiming that his theories of knowledge *cannot* be applied (at least to some extent) to programmed computers. It remains to actually apply them in this manner. I think the following quotation shows fairly clearly how this may be done:

> Because all our dispositions are in some sense adjustments to invariant or slowly changing environmental conditions, they can be described as *theory impregnated,* assuming a sufficiently wide sense of the term 'theory'. What I have in mind is that there is no observation which is not related to a set of typical situations—regularities—between which it tries to find a decision. And I think we can assert even more: *there is no sense organ in which anticipatory theories are not genetically incorporated.* The eye of a cat reacts in distinct ways to a number of typical situations for which there are mechanisms prepared and built into its structure: these correspond to the biologically most important situations between which it has to distinguish. Thus the disposition to distinguish between these situations is built into the sense organ, and with it the *theory that these, and only these, are the relevant situations for whose distinction the eye is to be used.*
>
> Popper (1970b, pp. 71–72, original emphasis)

While this quotation centers on subjective knowledge relating to sense organs, I take it that the general principles espoused here can be applied quite generally. Given this, I suggest that what Popper calls an "anticipatory theory" can be identified with what I have called a "predictive model". What he calls a "disposition" thus exists only in the context of an anticipatory theory (or predictive model), and can be identified with the contingent interaction of an anticipatory system (S_2) with the system of which it has a model (S_1). But these are precisely the conditions under which I have said that the model refers to the object modeled, or that the anticipatory system has *knowledge* of it. Thus, I claim that the knowledge which I propose to ascribe to (tokens of) computer programs is a *bona fide* case of subjective knowledge in at least one sense which Popper would allow or recognise.

While it is not crucial for my purposes here, I may say that Popper seems to distinguish this limited or impoverished kind of subjective knowledge (which, evidently, a computer *may* realise) from the more general kind (which, on Popper's view, a computer *cannot* realise) by reference to a hierarchy of *linguistic* functions. Specifically, Popper allows that both animals (including the human one) and machines can support two "lower" functions (the "expressive" and "signalling" functions), but he goes on to argue that there exist (at least) two further functions (the "descriptive" and "argumentative") which seem to be exclusively human achievements. Popper has discussed these ideas in a number of publications, but the clearest and most pertinent to the question of "machine" knowledge, is probably his paper on *Language and the Body-Mind Problem* (Popper 1953). In any case, the arguments for this hierarchy of language functions, and for the resulting cleavage of Popperian "subjective knowledge" into a kind that computers *can* realise and a kind that they *cannot* realise, seem to me to be largely equivalent to Popper's more general arguments against the causal closure of World 1. So, once again, I shall not pursue these questions further here.

This covers the application of subjective knowledge to computational systems. It leaves open the relevance, if any, of what Popper has called *objective* knowledge. Popper has generally identified objective knowledge with World 3, and I use the terms synonymously. As already noted, Popper has argued that this World 3 is not completely reducible to World 2. Still without reopening that debate, I want to emphasise that Popper accepts that in many cases World 3 objects can, in some sense, be identified with (if not reduced to?) World 2 or even World 1 objects. That is, objective knowledge can be *physically embodied.*

In general, in speaking of embodiments of World 3 objects, Popper has in mind *linguistically expressed* theories (which can then be subject to discussion and criticism). If this were the only case of the embodiment of objective knowledge, then it might have little immediate relevance to Artificial Intelligence—given the limited linguistic capabilities of existing computer systems. However, Popper also considers the concept of objective knowledge in a more general sense:

> ... all the important things we can say about an act of knowledge consist of pointing out the third-world objects of the act—a theory or proposition—and its relation to other third-world objects ...
>
> Popper (1970a, p. 163)

It seems to me that the implication here is that subjective *knowledge,* as such (as opposed to other World 2 objects, such as hopes and fears and pains etc.) should, quite generally, be viewed strictly in terms of its relationship to some World 3 object (the object being "grasped", in Popper's terms). In terms of my discussion of Anticipatory Systems then, I suggest that the predictive model *itself* (separately from the token which embodies or realises it) can be identified as an example of objective knowledge. That is, any systems (including computational ones) which can be ascribed subjective knowledge can also be said to *grasp,* even if only in a rudimentary fashion, some objective knowledge. They have, as it were, a toehold (at least) into World 3. This is a significant point, to the extent that Popper has stressed that, as a methodological guideline for scientific research, one should concentrate on the World 3, *objective* knowledge of a system, rather than its subjective World 2 realisation. He makes this kind of argument explicitly for non-linguistic (biological) organisms in (Popper 1968, p. 112–114).

In the present case, the significance of this is simply that, if we wish to ascribe meaning to a token (of a program) we can only do so by reference to the model which (we claim) it embodies. That is, the (tentative) identification of the World 3 object which a token embodies would be a crucial methodological step in any practical application of the computational semantics presented here.

That completes my review of computational semantics, or of the idea of artificial, computational, *knowledge.* In conclusion, then, let me just reiterate the central theme, which I have tried to view from a number of different perspectives. This is, firstly, that a programmed computer is a *dynamic* system, which must not be confused with the static, formal, object, which is its program; and secondly that such programmed computers can and should be said to be *knowledgable* precisely to the extent that they embody predictive model(s) of the reality in which they are embedded *and* that they *use* these predictions to condition their confrontation with that reality.

3.6 On the "Engineering" of Knowledge

At this point I have more or less identified the problem of Artificial Intelligence with the problem of *artificial knowledge*, and I have elaborated what I intend by that latter phrase in some detail. To be sure, for our computer to (say) pass the Turing Test, it must not only know about the world, but also be able to communicate (linguistically) about it; and to be sure, the specific problems associated with relating its knowledge to linguistic expression are far from trivial; but I suggest that the primary problem, in the current state of the AI art, is not that computers cannot talk, but rather that they have nothing worthwhile to say.

For example, as far as computer linguistic performance goes, we may consider Winograd's SHRDLU system to be a climax of sorts (Winograd 1973; Hofstadter 1979, pp. 627–632). The emphasis in the development of SHRDLU was on language "interpretation" or "understanding", rather than on language "production", but perhaps this is the harder of the two. In any case, viewed purely in terms of its ability to use language, SHRDLU was a considerable achievement—it could indeed maintain quite a creditable and coherent conversation.

Unfortunately, the conversation turns out to be extremely monotonous, or even boring. SHRDLU's "knowledge" of the world is limited to an extremely narrow and restricted domain, or a *microworld.* SHRDLU's particular microworld may be thought of as a table top with various kinds of toy-like blocks on it—cubes, pyramids, etc. in various colours. There is also an arm (belonging to, or operated by, SHRDLU) which can be used to move these objects around. I say that the microworld may be *thought of* in this manner, but, as always, one must be careful about who (or what) is doing the thinking here. SHRDLU's knowledge certainly encompasses some of the most salient aspects of the microworld I have described; but it also lacks all of the background ramifications that the description I have given would have for a human. Thus, not only is the *scope* of SHRDLU's knowledge very limited, but so also is the *depth.* It might be more accurate to describe SHRDLU's microworld as consisting of a 'table top', with 'blocks' 'on' it, of various 'shapes' and 'colors' etc.—using the scare quotes to emphasise that, although SHRDLU may use these terms in conversation, its understanding of them is, at best, a pale shadow of the normal human understanding of them.

So, it seems that the central problem is knowledge. It may, or may not, be a difficult problem to "hook up" an already knowledgable subject, so that it could communicate linguistically; but this problem hardly even arises until the system is quite knowledgable to start with: until it shares enough knowledge of the world with us that it might conceivably have something substantive to communicate. I emphasise this distinction between knowledge and the ability to linguistically communicate it, because there is sometimes a danger of confusing language and its content. This is closely related to the issue dealt with in the previous section, of the difference between a program *per se* (which knows nothing) and a programmed computer (which may or may not know something).

Now, the simplest conceptual approach to the problem of artificial knowledge is to *engineer* it. That is, one attempts to explicitly formulate model(s) of reality, and then instantiate them in a computer program; in other words, one builds an anticipatory system by actually designing and building the requisite predictive model(s) and using the output of these models in some (more or less) rational or appropriate way to condition the behaviour of the system—in particular, to condition its interaction with the object(s) modelled.

Knowledge Engineering is thus a brute force, or stipulative, approach to realising AI. It is the approach which has dominated AI research until relatively recently. It is (in effect) the way SHRDLU's knowledge was created, and is characteristic of AI's principle commercial success, the notion of the *Expert System.*

The question which now arises is: what is the scale of this (knowledge engineering) task? How much knowledge[15] does a typical human being have? Or, perhaps slightly less demandingly, how much knowledge would be required to pass (or even come close to passing) the Turing Test?

Turing himself attempted this kind of analysis. He first estimated the "storage capacity" of the brain at about 10^9 bits, and then comments:

> At my present rate of working I produce about a thousand digits of programme a day, so that about sixty workers, working steadily through the fifty years might accomplish the job, if nothing went into the waste-paper basket. Some more expeditious method seems desirable.
>
> Turing (1950, p. 455)

[15] I pretend, for the sake of the discussion, that there could be some meaningful quantitative measure of knowledge—say something like "person-years of development effort" to realise the corresponding artificial predictive model(s).

With the benefit of forty years experience of the problems of large scale software engineering, we might be permitted some wry amusement at Turing's even contemplating the idea of developing roughly 100 MByte of software, without anything going in the "waste-paper basket"; furthermore, Turing admits that 10^9 bits is a low estimate for the storage capacity of the brain. However, the point is that these factors only serve to further strengthen Turing's conclusion that, even supposing the knowledge engineering approach to be *theoretically* tractable, it is not *practical.*

While we might now be more reticent about doing this kind of calculation, nothing in the past forty years has served to suggest that Turing may have seriously *under*-estimated the effort required. That is, the knowledge engineering approach has proved more or less successful in narrow domains of knowledge, but it has remained limited to such domains. In terms of the original objective of general intelligence, at the Turing Test level, the approach has largely stagnated.[16]

The apparent limitations of knowledge engineering have been recently documented by Hubert and Stuart Dreyfus (Dreyfus & Dreyfus 1986). They identify two related difficulties: the *common sense knowledge* problem, and the *frame* problem.

The common sense knowledge problem refers to the extreme difficulty which has been encountered in attempts to systematise common sense knowledge. This is generally agreed to be a very severe problem, though there is room for debate as to its exact nature—specifically, whether it is "merely" a matter of scale, of the sheer quantity of knowledge involved, or whether there are more fundamental problems not yet properly recognised. Thus, for example, Hayes (1979) proposed a research programme to systematise or formalise "a large part of ordinary everyday knowledge of the physical world"—what he dubbed *naïve physics.* Drew McDermott was originally an enthusiastic advocate of Hayes' approach, but subsequently (McDermott 1987) reported that very little progress had been made, and concluded that the programme faced very fundamental and substantial difficulties.

[16]One major exception is Lenat's Cyc project (Lenat & Guha 1990). However, substantive results (one way or the other) are not expected from this project before about 1994.

The frame problem refers to the fact that, even if a system has been provided with a great deal of knowledge (without, for the moment, trying to quantify this), it is very difficult to integrate this successfully—especially to ensure that the most relevant knowledge is available and applied at any given time; and, of course, the more knowledge is provided, the worse this problem becomes. Dreyfus & Dreyfus describe the frame problem as follows:

> In general skilled human beings have in the areas of their expertise an understanding that enables them, as events unfold, to distinguish what is relevant from what is not. However, during the first three phases of AI research, from cognitive simulation up through work on micro-worlds, computers, like beginners, advanced beginners, and competent performers, were programmed to confront all facts as isolated from each other and goals as just further facts. Thus whenever a change occurred the whole set of facts that made up the computer's representation of the current state of affairs had to be recalculated to update what had changed and what had remained the same. The attempt to capture human, temporal, situated, continuously changing know-how in a computer as static, de-situated, discrete, knowing that has become known as the frame problem.
>
> Dreyfus & Dreyfus (1986, p. 82)

It is interesting to compare this with Popper:

> At every instant of our pre-scientific or scientific development we are living in the centre of what I usually call a *'horizon of expectations'*. By this I mean the sum total of our expectations, whether these are subconscious or conscious, or perhaps even explicitly stated in some language. Animals and babies have also their various and different horizons of expectations though no doubt on a lower level of consciousness than, say, a scientist whose horizon of expectations consists to a considerable extent of linguistically formulated theories or hypotheses.
>
> Popper (1949, p. 345)

My point here is that while Popper has never explicitly addressed the frame problem, it is clear that his theory of knowledge encompasses the issues it raises. To the extent that I have argued in the previous section that computers can, in principle at least, realise knowledge in Popper's sense, this can be taken as a claim that computers can, in principle, be programmed to overcome the frame problem. This is worth stating explicitly because Dreyfus & Dreyfus are frankly skeptical about it.

However, the point remains that, to date, the brute force method of knowledge engineering has proven to be extremely limited as an avenue toward the realisation of artificial intelligence. It seems that some alternative should be sought.

3.7 Building a Baby

Turing himself had, of course, anticipated that what I now call the knowledge engineering approach might prove impractical. His proposed alternative was to develop a machine which would be capable of *learning*. In this way, he hoped, the *initial* programming requirement could be reduced to manageable proportions:

> Presumably the child-brain is something like a note-book as one buys it from the stationers. Rather little mechanism and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child-brain that something like it can be easily programmed. The amount of work in the education we assume, as a first approximation, to be much the same as for the human child.
>
> Turing (1950, p. 456)

This possibility has certainly not been ignored in the intervening years. However, Turing's "hope" that it might prove significantly easier to program a "child-brain" compared to an adult has, so far at least, proved forlorn. As Charniak and McDermott put it:

> One idea that has fascinated the Western mind is that there is a general purpose learning mechanism that accounts for almost all of the state of an adult human being. According to this idea, people are born knowing very little, and absorb almost everything by way of this general learner. (Even a concept like "physical object," it has been proposed, is acquired by noticing that certain visual and tactile sensations come in stable bundles.) This idea is still powerfully attractive. It underlies much of behavioristic psychology. AI students often rediscover it, and propose to dispense with the study of reasoning and problem solving, and instead build a baby and let it just learn these things.
>
> We believe this idea is dead, killed off by research in AI (and linguistics, and other branches of "cognitive science"). What this research has revealed is that for an organism to learn anything, it must already know a lot. Learning begins with organized knowledge, which grows and becomes better organized. Without strong clues to what is to be learned, nothing will get learned.
>
> Charniak & McDermott (1985, pp. 609–610)

Popper has made related claims in a different context:

> ... to every man who has any feeling for biology it must be clear that most of our dispositions are inborn, either in the sense that we are born with them (for example, the dispositions to breathe, to swallow, and so on)

> or in the sense that in the process of maturation, the development of the disposition is elicited by the environment (for example, the disposition to learn a language).
>
> Popper (1970b, p. 66)

> If it were not absurd to make any estimate, I should say that 999 units out of 1,000 of the knowledge of an organism are inherited or inborn, and that one unit only consists of the modifications of this inborn knowledge...
>
> Popper (1970b, p. 71)

Boden (1988, p. 187–188) also discounts Turing's original programme for similar reasons. Somewhat more caustically, but making an essentially similar point, the biologists Reeke & Edelman have said:

> In fact, consideration of the magnitude of the problem with due modesty suggests that perception alone is hard enough to understand, without attempting to jump directly from perception to learning, through learning to social transmission and language, and from there to all the richness of ethology. At present, it is still a large challenge to understand how an animal can even move, and it would be well for AI to look first to such fundamental issues.
>
> Reeke & Edelman (1988, p. 144)

I note that this criticism by Reeke & Edelman is directed just as much at the recent revival of research in the field of artificial neural networks ("connectionism") as at the approaches of conventional (so called "symbolic") AI.

This result—that the realisation of an artificial "infant" intelligence seems not to be significantly easier than the realisation of an "adult" intelligence—is certainly disappointing, but it is by no means completely fatal to the AI enterprise. With my rejection, in the previous section, of the knowledge engineering approach the problem had already changed from that of artificial knowledge in itself, to the problem of the *growth* of (artificial) knowledge. That still remains our problem, but now we recognise that this cannot be solved by restricting our attention to the somatic time growth of human intelligence. We must take a more comprehensive view, in which human knowledge is seen as being continuous with animal or biological knowledge. In short, we must understand not only "learning" but also "evolution".

3.8 The Growth of Knowledge

The development so far has indicated that, in attempting to realise artificial intelligence, we must realise the *growth* of artificial knowledge; and that, furthermore, we must be concerned not just with the somatic time "learning" of an individual, but also with the evolutionary time growth of "inate" knowledge. That is, we must ultimately seek to realise the growth of artificial knowledge in something comparable to *both* of these biological senses.

I now wish to go beyond this result, and review a stronger, perhaps even a radical, claim: this is that the growth of knowledge by learning and by evolution are not fundamentally distinct processes in any case—rather, they are both forms of a kind of abstract or generalised Darwinian process. This being so, it will follow that the AI research programme may be identified with, or even replaced by, a programme aimed at the realisation of *Artificial Darwinism.*

3.8.1 Evolutionary Epistemology

The doctrine that the processes underlying *all* growth of knowledge are of an essentially Darwinian kind is now called *evolutionary epistemology*; the concept was pioneered by Popper, and it is fundamental to his overall philosophy, but it has also been significantly expanded and developed by others—Radnitzky & Bartley (1987) provide a comprehensive survey. I have very little to add to this existing literature, so I shall restrict myself here to a relatively brief review.

Evolutionary epistemology derives, ultimately, from Popper's analysis of the problem of *induction,* and the implications he draws from this for the growth of knowledge (e.g. Popper 1971). Popper denies that there can be such a thing as *certain* knowledge (except in the trivial sense of a tautology); and, more importantly, denies that there can be such a thing as a *logic* of induction. That is: nothing that we know is *necessarily* true (including "observation statements", since these are, themselves, theory impregnated); and even to the extent that what we know is, in fact, true, we cannot *logically* infer from it (consciously or otherwise) any more general, or strictly new, knowledge.

To take a favoured example of Popper's (e.g. Popper 1970b, p. 97) we know that the Sun rises each day; but this is not certain knowledge (there are any

number of reasons why the Sun may *not* in fact rise tomorrow); and our (tentative or conjectural) knowledge that the Sun will, in fact, rise tomorrow, is not, and cannot be, a consequence of, nor justified by, our experiences of the Sun rising on previous days—no matter (for example) how many times this experience may have been repeated.

It is important to emphasise here that, in this Popperian view, deduction as such can never result in *new* knowledge or in the growth of knowledge. Deduction is a tool which can be used (and most notably *is* used in science) to draw out consequences of our existing knowledge; but this is (just) making explicit what was already implicit, and, in itself, cannot increase our knowledge. Processes weaker than logical deduction (e.g. so called fuzzy logic) cannot, of course, reliably do any better in this particular respect.

Popper has called the naïve empiricist idea that knowledge is "derived" or "extracted" or "distilled" from some accumulation of "experience" the *bucket theory of knowledge* (Popper 1949). His central point is that, as long as knowledge is interpreted in the sense of effective predictive models, the bucket theory is untenable *on purely logical grounds.*

But if knowledge cannot grow through the analysis of experience, then how *does* it grow?[17]

Popper's answer is that knowledge grows, and *can* only grow, by a process of *conjecture and refutation.* By "conjecture" he means the formulation, *by any means*, of new models or assertions or theories, which make predictions about the environment in which the knowledge agent is embedded. The only constraint is that the predictions of these new models must potentially go beyond (or conflict with) predictions made by the prior knowledge of the agent.

In effect, we must divide knowledge processes into two kinds. In the first, truth conditions (or more strictly, *belief* conditions) are preserved. Such processes clearly do not involve any growth of knowledge, but rather represent an *elaboration* of existing knowledge; they include, for example, the operation or execution of "predictive models", as I have used the term (following Rosen 1985a)

[17] I discount here the relativist (non-)answer that the growth of knowledge is an illusion; whatever about the growth of human knowledge, the idea that the growth of biological knowledge, in the evolutionary sense, is imaginary, seems to me to be quite unsustainable.

in section 3.5 above. These processes correspond to the *application* of knowledge which has been accepted or adopted (at least tentatively) by an agent. In the second kind of knowledge process, truth (belief) conditions are not strictly preserved. Since these processes are not truth preserving, their output is inherently conjectural (i.e. even relative to premisses which are assumed to be true). Regardless of the nature of these processes, we may say that they represent *unjustified* variation. Such unjustified variation is clearly distinct from elaboration of already accepted knowledge—for it potentially produces conjectures transcending, and especially *contradicting,* the previously (presumed) known. But unjustified variation does not yet represent *growth* of knowledge, for these new conjectures may be uniformly mistaken.

Knowledge can *grow* if and only if the agent's predictions ("deductions") are *frustrated*; that is, if the world does *not* behave in the way the model (the knowledge) predicts; whenever expectations are defied in this way, there is an *opportunity* for growth. But this opportunity can be exploited only if the failure has the effect of selecting between competing models which were not all equally bad at predicting the behaviour of the world.

In short, there must be *some* mechanism for generating new candidate models of reality, which may compete with, and improve upon, the old ones; these may, or may not, be derived in some sense from existing models; but their validity or utility, is independent of their genesis. Knowledge can thus continue to grow only to the extent that new models of aspects of the world, not logically (deductively) entailed by prior knowledge, can be generated and tested. There can be no definite method (or logic of induction) for the generation of such new models which would guarantee their truth. For that matter, there cannot even be a definite method for *testing* of competing conjectures.[18]

While Popper originally formulated this theory of the growth of knowledge by conjecture and refutation in the context of *scientific knowledge*, the schema

[18]I may say that I fully accept that there is a concept being used implicitly here, of relative "closeness" to the truth, or "verisimilitude", which is a difficult and problematic one; but I think it has, nonetheless, a clear and useful intuitive meaning—in this, I follow Newton-Smith (1981), at least partially (compare also Popper 1974b, pp. 1011–1012). In any case, it must be emphasised that the notion in question is always a relative one: we are always talking about a comparison between competing conjectures rather than between an isolated conjecture and the "naked" truth (the latter being strictly inaccessible—to us just as much as to our machines).

clearly represents a kind of generalised or abstract *Darwinian* process. Campbell (1974a, p. 49) points out that this Darwinian undertone can be found even in Popper's earliest discussions of the subject; in any case, Popper himself has since (e.g. Popper 1961) explicitly emphasised the essential unity of all processes of knowledge growth in both evolutionary and somatic time (and also, indeed, in what we might call "cultural time", in the case of linguistically formulated World 3 knowledge; but that further extension is not relevant to my purposes here).

Thus, under the doctrine of evolutionary epistemology, all knowledge (subjective or objective, conscious or unconscious) consists in *tentative* hypotheses or theories about the world (including theories about theories, and how to apply them etc.). Growth of knowledge is possible (indeed, is *only* possible) through the unjustified formulation of new, tentative, theories, and some form of testing and selection between competing theories. In the special case that the theories are linguistically formulated, are falsifiable, and selection is based on rational analysis and critical testing, then Popper identifies this as *scientific* knowledge. In the special case that the theories are inate or inborn, and selection is based on differential reproduction arising through competition for limited resources, then the process is conventional Darwinian evolution. But, in all cases, the growth of knowledge involves an initial unjustified generation of new conjectures—i.e. conjectures whose truth is not logically entailed by the (tentatively accepted) truth of previous knowledge—followed by a confrontation between these new conjectures and the objective world, a confrontation in which the (more) false are rejected.

Campbell (1974b) has referred to this unified theory of knowledge growth as *Unjustified Variation and Selective Retention*, or, as I shall say, *UVSR*.

3.8.2 On "Random" Variation

The variation which underlies Darwinian (UVSR) evolutionary processes is commonly referred to as being "random", but it turns out that this has connotations which can be deeply misleading. Campbell (1974a; 1974b) has previously reviewed this question quite comprehensively; I shall simply extract some details which will be particularly relevant to my own objectives here.

Unjustified variation is an essentially *logical* notion. While it perhaps con-

forms to one of the common-sense ideas of "randomness", it is certainly *not* random in the sense of the probability calculus. This should be clear, for example, from the fact that the probability calculus relies on the possibility of a defined *event space*, whereas to classify a variation as "unjustified" does not *require* any reference to a "space" of "possible" variation. Lack of justification is, rather, a logical relationship between new (tentative) knowledge and prior knowledge.

More generally, "randomness" implies the absence of *predictability,* (except, perhaps, in a statistical sense) whereas "unjustified" variation may be arbitrarily systematic and predictable. Note that this does *not* imply that the growth of knowledge is predictable. As Popper (1988, pp. 62–67), for example, has pointed out, any claim to be able to predict the future growth of knowledge is fundamentally flawed; in the case at hand, such a claim must fail because the growth of knowledge requires both the generation of unjustified variations (and we stipulate that this *may* be predictable, in isolation) but also the testing, and selective retention, of some of the generated variations. This second step, of selective retention, is *not* predictable.

This establishes that unjustified variation is not (necessarily) a "random" process in the sense of the probability calculus, and that it need not even be unpredictable. But the crucial distinction between the notions of "unjustified" and "random" variation is rather more subtle than this. A key connotation of "random" variation, in the context of knowledge processes at least, is that it is "unbiased" with respect to the *truth*, or, more generally, the *verisimilitude*, of the generated conjectures. That is, a "random" variation should be "just as likely" to be *false* as to be *true.*

I should point out that, while this notion of randomness, in the sense of a lack of bias between truth and falsity, seems to have a fairly clear *intuitive* meaning, it can hardly be made formally respectable at all. To say that a process of generating conjectures is "random" in *this* sense requires that we be able to categorise all possible conjectures which it can generate as true or false *in advance*; and if we could do *that* then we would already have all the accessible knowledge in our possession, and there could be no growth. Granted, we could possibly apply this notion to our machines, or perhaps even to animals—if the domain of their knowledge is strictly circumscribed to lie within some domain of which

we already have "perfect" knowledge. But of course, we never have "perfect" knowledge of any domain, so even this is a contrived case; furthermore, machines whose knowledge cannot, *in principle*, transcend our own would still represent a critically impoverished kind of artificial intelligence.

In any case, the important point is that the idea of "unjustified" variation does *not* require or imply "randomness" in this last sense either. A particular process for generating new conjectures may, in fact, be strongly "biased" (either towards truth or falsity), insofar as this idea of bias can be given a clear meaning at all; but we can never *know* this to be the case. Our labelling of a generation process as *unjustified* does not rely one way or the other on such bias, or its absence.[19]

In fact, it seems that, if anything, "successful" UVSR processes typically involve generators which *are* strongly biased in favour of true (or "approximately true") conjectures. The crucial point here is that, while the *possibility* of knowledge growth does not rest on such bias, the *rate* of growth will be strongly affected by it.

In saying this it may appear that I am now begging the question at issue: it seems that I now explain or solve the problem of the growth of knowledge (at least insofar as it occurs at "speed") by calling upon some mysterious inbuilt bias in the generation of new conjectures. But this is actually a flawed criticism—because the UVSR principle can be applied recursively. If I should say (or conjecture) that a generator of (unjustified) conjectures is favourably biased, I would be implying that it already incorporates knowledge; while this would raise the question of where *this* knowledge came from, I have a simple answer ready—namely that it came from a prior process of (unjustified) generation of different *generators,* with selective retention of the "best". The implied regress is *not* vicious: it can bottom out with generators which are unbiased (or even unfavourably biased)—or, more to the point, generators whose operation can be explained without any

[19] I should caution that I myself have previously made the terminological blunder of using "unjustified variation" to refer not just to the relatively weak logical concept for which I use the term here, but *also* to refer to the stronger concept which I here call "unbiased variation" (McMullin 1992a; 1992b); this was a blunder insofar as I acquired the phrase "unjustified variation" from Campbell (1974b), and it has become clear to me that Campbell meant the term to imply *only* the weak, strictly logical, notion. I have therefore now reverted to using it *only* in this original sense of Campbell's.

assumption of bias (i.e. non-teleologically) one way or another.

So the "unjustified" in UVSR does not equal "unbiased" (or "ignorant"); but a complete evolutionary epistemology *does* demand that, to whatever extent an unjustified generator is held to already incorporate significant knowledge, that the genesis of *this* knowledge must be explained by a prior UVSR process. This recursion must ultimately terminate in "primordial" generators whose genesis is not problematic. We can hold that such a termination is *possible*, even if we have little or no idea of its detailed nature, because UVSR *can* operate even in the face of a strongly *un*-favourable bias (provided enough time is allowed).

This is an important result. The previous section established, on strictly logical grounds, that *unjustified* variation was necessary to the growth of knowledge—but said nothing about the *rate* of such growth. I am now suggesting that the rate of growth will depend on an ability to exploit previously gained knowledge in a loosely hierarchical fashion: that, in other words, for "fast" knowledge growth, we need an architecture which is not tied into a fixed, predefined, generator of unjustified variation, but which instead supports the emergence of new generators of unjustified variation, and selection between such generators.

There is a further separate, but complementary, result, regarding the "openness" of knowledge growth, but it will require some discussion to develop it properly.

Note firstly that, while it seems that an organisation supporting such a hierarchic knowledge structure may be a *necessary* condition for "fast" knowledge growth, it cannot, of course, be a *sufficient* condition. We may say that the fundamental principle of evolutionary epistemology is that there are *no* sufficient conditions for the growth of knowledge.

In practical terms, this means that if we happen to find a "good" generator of new (tentative) knowledge, then that can allow a burst of relatively rapid knowledge growth, but this will inevitably be exhausted; further growth of knowledge will then rely on generating an alternative generator. That is to say, a "good" generator is a double-edged sword: to the extent that it does generate good conjectures, it accelerates the growth of knowledge; but buried among all the "bad" conjectures which it does not generate, there may be some jewels, better even than the "best" conjectures which it does generate.

Thus, once it is accepted that *all* knowledge is conjectural—including that incorporated in our best "generators"—we see that the growth of knowledge may ultimately cease altogether if we cling dogmatically to *any* knowledge. Conversely, if we wish the growth of knowledge to be (as far as possible) open-ended, we need a knowledge structure which is *not* a simple hierarchy, but is rather more like a *heterarchy*, in which *all* knowledge (including all generators) is potentially subject to competition and displacement.[20] There is this inherent tension between the two aspects of the UVSR process—between variation and retention—and it is precisely the maintenance of this tension which permits (but cannot compel) continued, open-ended, growth of knowledge.

Our problem then is to find a way to design our putative machine or computer "intelligence" in just such a way that it can successfully balance on this same knife-edge which separates dogmatism from ignorance. This is not a trivial task.

3.8.3 UVSR and AI

The notion of realising some kind of more or less Darwinian process, in a computational system, is not at all original. Turing (1950) explicitly drew parallels between processes of learning and of biological evolution, in his seminal discussion of the prospects for realising AI. The earliest practical research was probably that of Friedberg and his colleagues in the field of "automatic programming" (Friedberg 1958; Friedberg *et al.* 1959). However, the problems they tackled were extremely simple (e.g. 1-bit binary addition), and the results mediocre. A form of artificial Darwinian evolution was proposed by Selfridge (1959), but was not pursued to an implementation. Simon (1969) provided an early, and perceptive, analysis of some of the general factors involved in applying any kind of Darwinian process in artificial systems.

The idea of artificial evolution was taken up again by Fogel *et al.* (1966), now specifically in the context of AI, but still applied to very simple problems and with very little tangible success. The scathing review of this particular work by Lindsay (1968) was, perhaps, instrumental in discouraging further investigation

[20] In its most abstract form this becomes, in effect, the principle of *pancritical rationalism* advocated by Bartley (1987). Compare also what I have previously called the "reflexive hypothesis" (McMullin 1990).

along these lines for some time. Holland has since rehabilitated the evolutionary approach somewhat with the so-called *Genetic Algorithm* or GA (Holland 1975), and this has generated a significant level of interest, particularly in the USA (Schaffer & Greffenstette 1988). I shall discuss the philosophical background to Holland's work in more detail below, and will also consider the Genetic Algorithm again in Chapter 4, section 4.3.2. There has been a somewhat parallel European development in the form of *Evolution Strategies* (Schwefel 1979; 1988). Several of these historical developments have been recently reviewed by Goldberg (1989).

I shall not attempt a comprehensive discussion of these prior efforts here. They have met, at best, with limited success, and then only in relatively narrow domains. I suggest that this failure can be traced, to a large extent, to the fact that these approaches have *not* been informed by the detailed philosophical arguments and analyses which have been elaborated by Popper and others under the rubric of *evolutionary epistemology.* I shall develop this claim by considering a number of attempts to probe the philosophical foundations of evolutionary or Darwinian growth of knowledge which have previously appeared in the AI literature.

I start with the analysis contained in Daniel Dennett's paper *Why the Law of Effect Will Not Go Away* (Dennett 1975). This is a remarkable paper in that Dennett succeeds in (re-)developing many of the important ideas present in evolutionary epistemology, but appears to have done so almost independently of, and concurrently with, the "mainstream" work in the field. There is only fleeting mention of Popper, with no detailed citation; and there is no mention at all of D.T. Campbell, or other workers associated with the development of evolutionary epistemology.

I note that Dennett's paper is clearly a development of ideas previously mooted in his *Content and Consciousness* (Dennett 1986), which was originally published in 1969. Thus this work by Dennett either predated or overlapped with the original publication of Popper's collection of essays *Objective Knowledge* in 1972 (Popper 1979), and the publication of the Schilpp volume on Popper's philosophy (Schilpp 1974), which also contained Campbell's most comprehensive expression of evolutionary epistemology (Campbell 1974a). So, notwithstanding the fact that the essential ideas of evolutionary epistemology were available in

much earlier publications[21] it is perhaps not too surprising that Dennett's development of these ideas was quite independent. While Dennett has mentioned Campbell's work more recently (Dennett 1981, p. 59), this was in a quite different context, and does not bear directly on the issues to be discussed here. As far as I am aware, neither Dennett himself, nor any other commentator, has previously drawn attention to the close connections between Dennett's analysis and evolutionary epistemology *per se*.

Although Dennett's treatment does not, in my view, go significantly beyond the analyses of Popper and Campbell, it is particularly relevant here because Dennett explicitly relates these ideas to AI.

Dennett expresses his discussion in terms of the thesis known, in behaviourist psychology, as the *Law of Effect*; very roughly, this states that actions followed by reward are likely to be repeated, or, more specifically, that rewards act to reinforce or *select* the "successful" behaviours from a diverse repertoire of possible behaviours. While Dennett is no friend of behaviourism, his claim is that there is a core of truth in the Law of Effect, and, in particular, that something like it will be "not just part of *a* possible explanation of behavior, but of *any* possible explanation of behavior" (Dennett 1975, p. 72, original emphasis).

Dennett develops this claim by first noting that Darwinian evolution provides an explanation of the growth of what I have called "inate knowledge" (and which Dennett refers to as purely "tropistic" or "instinctual" control of behavior); and that, for the time being at least, Darwinism is the *only* account we have of this growth which is not "question-begging". In Dennett's view, the Law of Effect provides a similarly unavoidable basis for any satisfactory account of the growth of knowledge in somatic time (i.e. what I termed "learning" above, as opposed to "evolution")—indeed he now reformulates the Law of Effect in a generalised form of what he calls "generate-and-test" procedures. These are hardly distinguishable from Campbell's UVSR processes—except that Dennett does not explicitly

[21] As already noted, there were clear anticipations of evolutionary epistemology in Popper's *Logik der Forschung*, first published in 1934, of which the English translation, *The Logic of Scientific Discovery*, first appeared in 1959 (Popper 1980); there were also substantive specific discussions of the problem of induction in Popper's *Conjectures and Refutations* (Popper 1989), first published in 1963; and Campbell published two seminal papers in the field at an early date (Campbell 1960a; 1960b).

formulate the notion of logically *unjustified* variation as a essential element of these processes.

Dennett's analysis closely mirrors that of Campbell in other respects also. In particular, he introduces the notion of an "inner" environment which can allow a selection process to proceed internally to an organism, apart from overt external behaviour. This is very similar to Campbell's idea of "vicarious" or "substitute" selectors (Campbell 1974a), and, like Dennett, Campbell has explicitly viewed this as a necessary generalisation of earlier learning theories, behaviourist and otherwise (Campbell 1960a). Following Simon (1969, Chapter 4, pp. 95–97), Dennett's discussion of the rôle of generate-and-test procedures in AI focuses on the requirement (if the growth of knowledge is to be "efficient") for generation processes to be "endowed with a high degree of selectivity" (Dennett 1975, p. 86)—which I take to be equivalent to what I have earlier called "favourable bias". However, unlike Simon, but exactly matching Campbell's various descriptions, Dennett explicitly recognises that any such "selectivity" in a generation process itself demands an explanation in turn, and that this can only be satisfied by a recursive appeal to some earlier (perhaps properly Darwinian or evolutionary) generate-and-test process; and, further, that this recursion can only bottom out with an appeal to some "ultimate" generators which "contain an element of randomness or arbitrariness" (Dennett 1975, pp. 86–87).

So: it seems clear that Dennett's ideas, so far at least, are essentially at one with the ideas of evolutionary epistemology. But I have not yet dealt with Dennett's substantive claim: that his generalised Darwinism, in the form of generate-and-test, is a *necessary* component of any satisfactory psychology (or, indeed, AI). Frankly, I find Dennett's argument here obscure, and I shall not attempt to reproduce it. But I note that, in introducing his argument, Dennett says:

> I suspect this argument could be made to appear more rigorous (while also, perhaps, being revealed to be entirely unoriginal) by recasting it into the technical vocabulary of some version of "information theory" or "theory of self-organizing systems". I would be interested to learn that this was so, but am content to let the argument, which is as intuitive as it is sketchy, rest on its own merits in the meantime.
>
> Dennett (1975, p. 84)

Now I admit that I may not have grasped Dennett's argument correctly; but insofar as I think I do understand it, it seems to me that it is, indeed, "entirely

unoriginal" (as Dennett anticipated) and that it already *had* been made more "rigorous". However, far from being related to "information theory", I think it is actually, in essence, a specialised form of Popper's argument in relation to the impossibility of a logic of induction. It differs primarily in that Dennett has failed to locate the problem as precisely or clearly as Popper, as being that of the growth of knowledge (in the face of the impossibility of induction), and his solution is, as a result, much less clearcut than Popper's; but it *is* essentially the same result, namely that there is no logical alternative to seeing knowledge as irredeemably hypothetical, and that knowledge grows, if at all, by (unjustified) variation and selective retention.

Having accepted that Dennett's analysis of the growth of knowledge is essentially equivalent to the doctrine of evolutionary epistemology, the remaining question is the relevance of this to AI.

Dennett notes that generate-and-test, in the most general sense, is a "ubiquitous" strategy within AI programs (and actually uses this fact to bolster somewhat his argument for the *necessity* of such processes). That is fair enough, as far as it goes, but it does not go very far. The important point for my purposes is that the use of generate-and-test is not, *in itself,* any kind of panacea. Dennett argues (and I agree) that such processes will be necessary in the realisation of AI—but they are not sufficient by any means. I have argued earlier that the effective and open-ended growth of knowledge actually requires that the architecture of the knowledge agent must support the growth and elaboration of a heterarchical structure in which no knowledge (including knowledge generators, and vicarious selectors) is sacrosanct, or dogmatically held. This is a much stronger requirement than simply insisting on having "some" form of UVSR; as far as I am aware, such an architecture is unknown in any functioning AI system.

In particular, Dennett himself (who is, of course, primarily a philosopher) has not attempted to actually apply the abstract ideas reviewed here in the design of any real AI system, and I shall therefore not discuss his analysis any further.

I now turn to some writers who, at first sight at least, might seem to oppose the central claim that UVSR processes are essential to the growth of knowledge, and thus to the realisation of AI.

Consider first the criticism by Boden (1984). Boden is primarily concerned

with the relevance of Darwinian processes to problems of mentality, specifically including AI, but she is also prepared to carry her attack to the home ground of Darwinian theory, biological evolution itself:

> Perhaps similar considerations concerning creative exploration might illuminate various biological phenomena which, on a neo-Darwinist account of evolution, are very puzzling. These include the facts that the fraction of DNA that does not code for the synthesis of specific proteins increases phylogenetically; that species have evolved remarkably quickly, and that the more complex species have if anything evolved at a greater rate than their predecessors; and that the speed with which a species evolves morphologically seems quite unrelated to the rate at which its individual proteins evolve (so frogs have protein-synthesizing mechanisms of comparable complexity to those of man). Such facts are not explicable in terms of "Random-Generate-and-Test," the mutational strategy favoured by neo-Darwinism. This is because (as was discovered by the early workers in automatic programming), the combinatorics of such a process are horrendous (cf. [Arbib 1969a[22]]). Switching to a higher-level biological language (cf. "consolidation"), might be effected by random processes of gene duplication and recombination; but this merely reduces the exponent without preventing an exponential explosion.
>
> Instead, some strategy of "Plausible-Generate-and-Test" is needed, whereby *mutations of a type likely to be adaptive become increasingly probable.*
>
> Boden (1984, p. 312, emphasis added)

But although Boden represents herself here as being opposed to "neo-Darwinism", it should be clear that there is, in fact, very little difference between the position she describes and the general position envisaged within the scope of evolutionary epistemology. Specifically, Boden seems to be assuming that neo-Darwinism must rely exclusively on generators of variation which are *unbiased*; but as I have already explained, that is a mistaken view.[23] The structure of Darwinian explanation (or, more generally, of evolutionary epistemology) demands only that, to whatever extent a generator of variation exhibits significant favourable bias, this will require a further, recursive, invocation of the UVSR principle; and this can bottom out only with generators whose explanation or genesis is, we may say, unproblematic.

---

[22]I have been unable to identify the relevance of Boden's citation here: (Arbib 1969a) makes no reference, that I can see, to automatic programming; it seems possible that the intended reference was actually to (Lenat 1983).

[23]While I believe that Boden is mistaken in her interpretation of neo-Darwinism, I also consider that she can hardly be blamed for this. Evolutionary biologists have not always been very clear on the issue, though there have been some useful recent discussions (e.g. Dawkins 1989a; Wills 1991). I review this in more detail in (McMullin 1992b).

Boden evidently accepts all this: thus she says, variously that "the initial heuristics must evolve by random mutation (since there is no suggestion of teleology here)" (p. 312), and that "a structural theory can even allow that contingency is sometimes *essential* to creative intelligence" (p. 314).

Nonetheless, a literal reading of the full passage quoted above might still suggest that Boden has something stronger than this result in mind: the last sentence, in particular, with its reference to favourable mutations becoming "increasingly probable", could be read as implying some kind of *inevitable* progression, or even acceleration, in the growth of "adaptation" (or, in my terms, in the growth of knowledge). That would obviously be deeply contrary to the principles of evolutionary epistemology. But, I doubt that Boden herself really intends such a strong claim, for there is no explicit *argument* to such an effect. I shall not, therefore, consider her analysis any further.

By contrast, I think that Lenat (1983—a work which is heavily referenced in Boden's discussion) *does* have a genuine, if implicit, disagreement with the principles of evolutionary epistemology; but, equally, I believe that Lenat is wholly mistaken in this. Lenat's epistemology seems, at that time at least, to have been a naïve inductivism:

> The necessary machinery for learning from experience is not very complex: accumulate a corpus of empirical data and make simple inductive generalizations from it.
>
> Lenat (1983, p.287)

Lenat was evidently unaware that the idea of induction presented any difficulties *in principle.*

I should note that Lenat was working on a system (`EURISKO`) which went some way toward meeting the architectural requirements I have identified for realising AI. Specifically, `EURISKO` included components ("heuristics") for generating new conjectures, and the system was *reflexive* in the sense that these heuristics could operate on each other. However, the heuristics seem to have been relentlessly inductivist, and `EURISKO` cannot be viewed as implementing UVSR in any reasonable sense. In any case, the system had very limited success, and Lenat himself subsequently abandoned this line of research; but he has not, apparently, abandoned an essentially inductivist epistemology. In a recent discussion of the

on-going Cyc project, he outlines the following objective for the system (to be achieved by late 1994):

> **Demonstrate that Cyc can learn by discovery. This ... includes deciding what data to gather, noticing patterns and regularities in the data, and drawing from those patterns useful new analogies, dependencies, and generalizations.**
>
> **Lenat & Guha (1990, p. 357)**

Thus, given that Lenat has still not recognised, much less analysed, the philosophical problems underlying induction, I suggest that any criticism of evolutionary epistemology implied in his work can be safely neglected.

To close this review of philosophical work bearing on the relation between UVSR and AI, I shall consider the position of John Holland and certain of his co-workers.

Holland is an engineer and scientist who has long been concerned with the problems of developing artificial "adaptive systems", and was in the vanguard of those advocating an evolutionary approach to such problems (e.g. Holland 1962b; 1962a). In particular, as noted earlier, Holland is the inventor of the so-called *Genetic Algorithm* (Holland 1975), a general purpose "learning" or adaptive procedure inspired in certain ways by the mechanisms of biological evolution. Holland has specifically attempted to apply the Genetic Algorithm in the development of machine learning systems which could overcome the brittleness of conventional expert systems (Holland 1986). It might seem, therefore, that Holland would surely support the position I have been advocating—that reflexive, heterarchical, UVSR processes are essential to the efficient, open ended, growth of knowledge. It transpires, however, that that would be, at best, an oversimplification.

In originally introducing the Genetic Algorithm, Holland identified himself as being concerned with the growth of "adaptation" (Holland 1975, Chapter 1). More recently, Holland has written explicitly in terms of the growth of "knowledge", particularly in the volume (Holland *et al.* 1986), co-written with K.J. Holyoak, R.E. Nisbett and P.R. Thagard; here they jointly identify themselves as concerned with "all inferential processes that expand knowledge in the face of uncertainty" (p. 1).

However, the situation is rather more complicated than this. The philosophical framework underlying Holland's approach is, at first sight at least, quite

incompatible with that which I have been advocating: like Lenat, Holland *et al.* seem to be self-declared *inductivists*—indeed, the title of their book (Holland *et al.* 1986) is actually *Induction*. Despite this, I shall *not* be arguing against the epistemology, as such, of Holland *et al.*; instead, I shall suggest that the appearance of disagreement is mistaken, a matter of words and emphasis rather than substance. But, in the very process of reconciling these apparent differences, I shall conclude that the aspects of the growth of knowledge which Holland *et al.* choose to concentrate upon are largely those which I choose to neglect, and *vice versa*.

The *appearance* of disagreement between the position of Holland *et al.* and that of Popper and Campbell is clear enough. We find, for example, an unqualified rejection of "evolutionary epistemology" (Holland *et al.* 1986, p. 79; they explicitly cite Campbell 1974a; 1974b); and while we (eventually) find an admission, mentioning Hume and Popper, that the very possibility of induction is problematic, this is immediately passed over with the statement that "most [philosophers] have attempted to solve the narrower problem of determining under what circumstances the [inductive] inference can be justified" (Holland *et al.* 1986, p. 230).

However, I do not think the situation is quite as it may seem when selectively quoted in this way. While the writers apparently *believe* themselves to be opposed to Popperian epistemology, their genuine familiarity with Popper's work can reasonably be questioned. There is only one explicit citation of Popper (on p. 328, to the first English edition of *The Logic of Scientific Discovery*—see Popper 1980). This is followed, quite shortly, by an ascription to Popper of the view that any reluctance to abandon functioning (but refuted) scientific theories must represent "an irrational, egotistical attachment" (p. 332); Holland *et al.* go on to suggest (apparently as a contrast to this "Popperian" view) that, in practice, theories can only be discarded "when a new, alternative theory that does not require special assumptions arises and offers to replace the existing theory" (p. 332). But one could hardly find a more genuinely Popperian statement that the latter: it is, precisely, the notion of survival between competing hypotheses. Popper (1974b, p. 995) has had occasion to defend himself against a similar "criticism", where he gives a more detailed rebuttal. For my purposes it is sufficient to note that the

rejection of a "Popperian" epistemology by Holland *et al.* may be more imagined (on their part) than real.

In fact, it seems that the "inductive" processes with which Holland *et al.* are concerned may be more or less identified with the "plausible-generate-and-test" processes of Boden (1984). Thus, in introducing their problem, Holland *et al.* quote C.S. Peirce at length to the effect that the growth of knowledge involves something like "special aptitudes for guessing right" (p. 4); and later (p. 79) they explicitly refer to this as "Peirce's problem of generating *plausible* new rules", which is almost identical to Boden's formulation (though the latter is not actually cited). This last reference to Peirce is the more ironic since it is immediately juxtaposed with the already mentioned dismissal of (Campbell 1974a)—a paper in which (pp. 438–440) Campbell carefully and critically reviews Peirce's profound ambivalence on the issues at hand.[24]

So: my conclusion here is essentially the same as previously outlined in reviewing Boden's work. I hold, essentially, that the processes which Holland *et al.* describe as *inductive* are processes of unjustified variation. Notwithstanding my use of the term "unjustified" here, I quite accept that, in given circumstances, some such processes may do "better" than others (in the sense of generating conjectures which are "closer" to the truth). Similarly, I accept that the formulation and comparison of processes in this respect is a genuine and difficult problem. But, crucially, I hold that there can be no final or definitive "solution" to this problem; all "inductive" processes are heuristic and fallible; there is no "logic" of induction. I say this without doubting, for a moment, that even partial solutions to this "problem of induction" may be very interesting, and, indeed, pragmatically useful. It seems to me that Holland *et al.* do not ultimately disagree with any of this, and that their analysis need not, therefore, be considered any further here.

[24]I might also add that Popper himself has made very positive remarks regarding the philosophy of C.S. Peirce (e.g. Popper 1965, pp. 212–213); however, these remarks do not bear directly on the issues under discussion here.

3.9 Conclusion

In this chapter I have argued that computational systems *can* be said to have knowledge, in a perfectly conventional, biological, sense; that this knowledge can grow *only* via some kind of Darwinian, UVSR, processes; that such processes will therefore be an essential component of any system pretending to human-like intelligence (as represented, for example, by Turing Test performance); and that, in any case, such processes (as opposed to pure "Knowledge Engineering") may well be essential to the *initial* construction of any system which exhibits, or aspires to exhibit, human-like intelligence.

It follows that the realisation, in a computational system, of UVSR processes incorporating an open-ended, reflexive, heterarchical, architecture—which is to say, in effect, some form of *Artificial Darwinism*—is now seen as being at least *an* essential element (if not *the* essential element) of a serious AI research program. The next chapter will be devoted to reviewing some issues which arise in practical attempts to do this. I shall leave the final summarising word for the present chapter with Popper:

> I do not really believe that we shall succeed in creating life artificially; but after having reached the moon and landed a spaceship or two on Mars, I realize that this disbelief of mine means very little. But computers are totally different from brains, whose function is not primarily to compute but to guide and balance an organism and help it to stay alive. It is for this reason that the first step of nature toward an intelligent mind was the creation of life, and I think that should we artificially create an intelligent mind, we would have to follow the same path.
>
> Popper & Eccles (1977, Chapter P5, p. 208)

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 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*LM*'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*CM* 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 UCM, (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:

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

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:

$$\forall\, m \in M_u,$$
$$\forall\, x \in X,$$
$$((u_0 \oplus x) \oplus d(m)) \rightsquigarrow m$$

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 U*G*M, 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:

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

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:

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

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 considerable emphasis on this point, referring to it as *intrinsic par-*

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.

Chapter 5

Artificial Genesis

5.1 Introduction

> ... AI as a field is starving for a few carefully documented failures. Anyone can think of several theses that could be improved stylistically and substantively by being rephrased as reports on failures. I can learn more by just being told why a technique won't work than by being made to read between the lines.
>
> McDermott (1976, p. 159)

I have argued that the central outstanding problem in the realisation of a substantive growth of A-knowledge via a process of Artificial Darwinism is that of exhibiting A-machines which are not only self-reproducing but also *robust*, in the face of a "hostile" environment. By "self-reproducing" I mean, of course, the von Neumann sense of supporting "heritable mutation"; that is, our A-reproducer should be a member of a set of A-reproducers, which span an indefinitely large range of A-complexity, where this set is connected under some form of A-mutation. The von Neumann schema of *genetic* self-reproduction shows, in outline at least, how this condition can be satisfied.

This outstanding problem, which I have labelled P_a, is still very informal; nonetheless, I have suggested that, to date, there has been little tangible progress toward a solution. Without attempting to prejudge the ultimate prospects for solving P_a by the "direct" route (i.e. by directly designing robust A-reproducers within some "reasonable" A-system) I have pointed out that there may be an alternative "indirect" approach—namely attempting to exhibit the *spontaneous* emergence of (viable) A-reproducers. The latter approach is inspired by the

(conjectured) spontaneous genesis of life in the biological world.

This chapter is concerned with a critical investigation of a class of A-systems which (it has been suggested) might indeed exhibit something like the spontaneous origin of Artificial Life.[1] I should warn in advance that the results to be presented here are largely negative: it will turn out that, contrary to expectations, the single specific A-system which will be examined in detail *cannot* support phenomena of this sort. However, I shall argue that the mechanisms of failure are not without interest.

The A-system presented here is an example of an α-Universe. The α-Universes are a class of artificial system originally proposed by Holland (1976). Holland made this proposal in a particular context, related to, but by no means identical with, my P_a. It is therefore useful to briefly review the problem situation which Holland *intended* to address with the α-Universes.

Holland's stated objective was to rebut certain criticisms of the neo-Darwinian interpretation of evolution. The situation was roughly as follows (for a more detailed discussion Holland cites Moorhead & Kaplan 1967):

- Darwinian evolution is predicated on the prior existence of entities having a wide behavioural repertoire which includes, among other things, the ability to self-reproduce in a manner which supports heritable mutation. Following Gould, I have previously called such entities Darwinian actors, or simply *D-actors* (Gould 1982; McMullin 1992a).

- Biological Darwinism should therefore be accompanied by some complementary theory to explain the advent of the *initial* D-actors in the so-called "primordial soup". I stress that the problem being presented here is *not* that of Biological Darwinism itself—i.e. whether the latter provides an adequate theory of the growth of biological complexity *once an initial set of biological D-actors is postulated.* It is the prior problem of whether Darwinian processes could have (spontaneously) started in the first place: the problem of the original genesis of life.

[1] A preliminary version of some of the material presented in this chapter has been previously published (McMullin 1992d). However, the treatment given here is much more detailed and extensive, and includes more recent experimental results: it may be regarded as the definitive version.

- This complementary theory should not draw on any new causal principles, over and above those assumed by biological Darwinism in the first place (it should not, for example, be theistic)—for otherwise, Biological Darwinism itself would evidently be undermined.

- A first (and naïve) tentative solution is the conjecture that, prior to the emergence of the initial D-actors, conventional physical effects (thermal and electrical agitation of the unorganised chemical soup), will result in the generation of a wide variety of physically feasible structures in an "unbiased" manner (i.e. we do not suppose that D-actors are any more likely to be generated than other structures of comparable "size"). Provided sufficient time is available, this process might eventually result (with probability approaching one?) in the emergence of the required, initial, D-actors. Darwinian evolution then takes over (or not, as the case may be; that is a separate problem).

- This conjecture seems, however, to be refuted by quantitative calculations of the expected time to emergence of D-actors, based on such an unbiased search process.[2] Even allowing for a substantial margin of uncertainty in the parameters for these calculations, the result is an emergence time so large that it seems entirely incompatible with such emergence having occurred in the lifetime of planet Earth.

- The next proposed solution is to retain the conjecture that conventional physical effects will result in the generation of a wide variety of structures, but to suppose that this generation process may be (or may become) strongly *biased.* In particular, in analogy with conventional Darwinian theory, it is conjectured that there may be *incremental* progress toward fully qualified D-actors. That is, there may be structures, typically much simpler than the ultimate D-actors (i.e. much simpler than even the simplest of contemporary *organisms*), which thus have plausible emergence times, but which might still have a long term effect on the subsequent generation

[2]This is a variant of the infamous "monkeys typing Shakespeare" kind of argument (e.g. Dawkins 1986, pp. 46–49).

process—biasing it in such a way that fully qualified D-actors can, feasibly, have emerged within the available time. To put it another way, we abandon any notion of a strict, binary distinction between structures which are "D-actors" and those which are not, and accept that there may be a continuum. Instead of supposing that there was a more or less dramatic or catastrophic change between unbiased generation of structures, and Darwinian evolution of structures, we suppose that what we now recognise as Darwinian evolution may have emerged gradually, reinforcing itself as it became established. We might call this a *bootstrap* theory of the emergence of D-actors (and Darwinian evolution).

- This serves to rescue a materialist theory of the origin of life and thus, of Biological Darwinism itself; but at the cost of becoming vague, qualitative, and, in this form, virtually untestable. In Popperian terms, it has become, not so much a theory, as a *metaphysical research program* (e.g. Popper 1976, Section 33, pp. 148–151). That is, it is a framework for the development of detailed theories, which detailed theories *might* then be capable of making testable predictions.

- It is extremely difficult to improve on this situation. Ideally, we would analyse and/or simulate or duplicate a quantitative model of the dynamics of the postulated primordial "soup" and thereby formulate a detailed, quantitative, testable, theory of the emergence of fully qualified D-actors. Indeed, considerable effort has been expended along these lines, with some degree of success (e.g. Oparin 1953; Eigen & Schuster 1979; Dyson 1985); however, these efforts are seriously limited by the size and complexity of the system (the primordial soup) being investigated. Thus: the detailed composition of the system is quite uncertain; the basic chemical interactions are varied, complex, and non-linear; the system is extremely large (say the total number of active chemical components on the planet Earth, during the pre-biotic epoch), and the duration potentially available for the significant processes is extremely long.

- Holland proposed an alternative approach (still within the same basic metaphysical research program). This is to investigate the behaviour of very much simplified systems, in order to provide a "proof-of-principle". The idea is this: suppose we can formulate a relatively simple system, which is, nonetheless, capable of demonstrating analogous phenomena to those being postulated for the primordial soup. Specifically, the system should be such that some form of D-actors can be sustained in the system, if they once emerge; that there is some kind of continuum of behaviour from that of simpler structures to that of fully qualified D-actors; and that, in the absence of bias due to the behaviours of the structures already present, there will be an unbiased generation of structures over some large set of "feasible" structures. We can calculate the naïve emergence time for D-actors, in an analogous manner to that for the "real" primordial soup; but additionally, if the system is sufficiently simple, we may be able to demonstrate analytically and/or experimentally, that D-actors can actually emerge in a much shorter time, by virtue of the generation process becoming progressively biased. If this could be achieved, it would constitute a proof of the principle that D-actors, and Darwinian evolution, could establish themselves spontaneously. This would not, of itself, "prove", or even "verify", the theory that some analogous process occurred in the real primordial soup; but it would increase our *preference* for such theories, by refuting the implicit alternative (that such behaviours are entirely impossible).

- To this end, Holland introduced the α-Universes as a class of simple A-system which could exhibit at least some of the required properties; he went on to identify one particular α-Universe (which I shall denote α_0) for which he was able to present detailed, closed form, analytic results. On Holland's analysis, α_0 demonstrates precisely the result sought: simple D-actors appear to have an expected emergence time much shorter than would be predicted by a naïve assumption of unbiased generation.

At first sight, Holland's analysis (if it survives critical testing) would solve not just the problem he was directly addressing, but P_a also: for, as I have described the situation, Holland seems to claim, *inter alia,* that α_0 can support D-actors—

i.e. *identified*, robust, A-reproducers. The fact that such D-actors should emerge spontaneously would, in that scenario, be an added bonus, but would be quite inessential to the solution of P_a.

However, on closer analysis the situation proves to be rather more complicated than this. While Holland does identify putatively robust, "self-replicating", A-machines, which can be embedded in α_0, these are not properly self-reproducing in the von Neumann sense; although they do involve a von Neumann style genetic mechanism, the complete set of related A-reproducers is essentially trivial—it certainly does *not* span a wide range of A-complexity.[3] Thus, α_0 certainly cannot offer an *immediate* solution to P_a.

The real relevance of α_0 to P_a is the following: if Holland's analysis of α_0 is correct, then it suggests that some more "powerful" α-Universe might support a set of D-actors (A-reproducers) spanning a satisfactory range of A-complexity, *while still retaining the property that such D-actors would spontaneously emerge.* If this were so, it might allow, as previously anticipated, an experimental solution of P_a *without* the need for an *a priori* design of any initial, robust, A-reproducers.

The question which immediately arises is whether Holland's analysis of α_0 is, in fact, correct. Although α_0 is extremely simple compared to real chemical systems, its analysis is by no means trivial, and could conceivably be mistaken. Holland therefore noted that his analysis could feasibly be tested by instantiating an α-Universe in a suitable, high speed, digital computer. However, Holland himself did not report on such tests, and, as far as I am aware, no such test program was ever carried out (Holland, Langton, personal communication).[4]

This chapter will therefore present original results from just such a program. I preface this with a more detailed and formal definition of α_0 than that originally presented by Holland, and an account of relevant aspects of the particular implementation.

As already indicated, these results will be negative: it transpires that Holland's analysis was indeed mistaken (through being oversimplified). In fact, even

[3]In this respect, the A-reproducers in α_0 are essentially similar to an A-reproducer proposed by Langton (1984) (in a rather different A-system); see my previous discussion in Chapter 4, section 4.2.7.

[4]Indeed, I have been able to identify only two substantive discussions of any kind of (Holland 1976). I shall return to these in section 5.5.8 below.

the extremely impoverished D-actors proposed by Holland for α_0 prove *not* to be robust; the question of their spontaneous emergence (never mind the emergence of more powerful D-actors in some alternative α-Universe) is thereby rendered irrelevant. This outcome will serve principally to reiterate again the seemingly intractable nature of P_a; but it will nonetheless also suggest some useful new insights into the problem.

5.2 The Universe α_0

5.2.1 Outline

Firstly, let me note that α_0 is not strictly a single, unique, α-Universe but denotes instead a parameterised family of related α-Universes. I shall identify such parameters as they arise, but otherwise it will be convenient to continue to refer to α_0 in the singular.

I should emphasise that Holland's original definition of α_0 was not complete; that is, many detailed aspects of its operation were left unspecified. The implication is that these details should not affect the ultimate outcome; nonetheless, in any realisation of α_0 it is still necessary to fill in all such details in some particular way. This section thus serves both to re-present Holland's original definition and also to specify, in detail, how this definition was extended and completed to allow a practical realisation.

Loosely speaking, α_0 consists of some fixed number of discrete *atoms.*[5] The total number of atoms is a parameter of α_0, and is denoted R; in general, Holland does not stipulate a specific size in his analysis. While he does refer to what he calls a "region" he gives no precise definition of "region". The experimental work described below will be based on a total size of $R = 10^4$ atoms, this being the size of "region" used by Holland for numerical calculations.

Atoms are classified into six distinct kinds, or *elements.* Of these, one has an especially distinguished rôle, and is referred to as the *null* element; the remaining

[5]Holland strictly speaks in terms of an underlying "cellular automaton", each cell of which can effectively "contain" one atom. The cells then remain "fixed" while the atoms "move" among the cells. However, this underlying cellular automaton *per se* plays no rôle either in Holland's analysis or the implementation to be described here; further discussion of it is therefore omitted.

five elements are collectively referred to as *material* elements. α_0 supports a detailed principle of "matter" conservation, in that atoms cannot be transmuted from one element to another, and thus the numbers of atoms of each element remain constant. The "densities" of each element (the number of atoms of the element divided by the total number of atoms R) are thus further parameters of α_0.

Each atom has, associated with it, one *bond*, connecting it to one other atom. A bond may be in either of two states: *strong* or *weak*. This state may be dynamically altered under the action of certain α_0 operators (with the exception that a bond originating with a null atom, or connecting to a null atom, cannot become strong).

As long as a given bond is strong, it cannot be disconnected. By contrast, a weak bond, may, in certain circumstances, be disconnected and re-connected to a different atom; in this way the connections between atoms may change in time. However, it is characteristic of α_0 that bonds may only be "transiently" disconnected; that is, every operator which involves disconnecting a bond also involves re-connecting it (to a different atom) all within a single time step. It follows that, before and after the operation of any allowed α_0 operator, every atom must be connected to precisely one other atom, and, therefore, that all the atoms in α_0 must form a single connected chain. Assuming that the number of atoms (R) is finite (as it must be for any practical realisation) this further implies that the chain of atoms must be closed on itself—i.e. it must form a single closed *loop*.

Any arbitrary connected series or sequence of atoms in α_0 will be called a *segment*. In effect, all α_0 operators which change the relative ordering of the atoms will do so in two stages (both completed in a single time step): a segment (which may consist of a single atom) is first cut out of one part of the loop, thus *transiently* dividing α_0 into a (smaller) loop and a separate, disconnected segment; this disconnected segment is then spliced back in, at some other point, reforming α_0 into a single closed loop of R atoms again.

A segment consisting of a null atom, followed by one or more material atoms, and then terminating in another null atom, is called a *structure*; a structure

containing exactly one material atom will also sometimes be referred to as a *free atom.*

A *complex* is a set of interacting structures: it is (for the time being at least) the kind of entity which we shall recognise as an *A-machine* in α_0. The structures making up a complex need not, in general, have definite connections with each other (a complex is not a segment *per se*); however, for a complex to exhibit interesting properties it is generally necessary that all the component structures be more or less "close" to each other.

The α_0 dynamics progress in discrete time steps; that is, the operators are all defined in terms of their effect in a single such time step.

Holland anticipated that, for a computer realisation of α_0, a time step might be accomplished in about 1ms of real time; however, he gave no indication of the kind of platform he assumed to achieve this. In any case, in the experiments to be described below, a more typical value actually achieved was of the order of 500 ms per time step (based on an Intel 80386 CPU with 33 MHz clock)—though this varied very considerably with the actual state of the universe.

The dynamic behaviour of α_0 is stochastic, and is defined in terms of two groups of operators: the *primitive* operators, and the *emergent* operators. The primitive operators are context *insensitive*—i.e. they apply throughout α_0 without regard to its sequential configuration. They are the abstract counterparts of diffusion and activation in real chemical systems. The emergent operators are context *sensitive*—i.e. their operation is sensitive to the sequential configuration of α_0. In effect, certain structures (should they arise) have special dynamic properties. They are termed "emergent" operators precisely because they are contingent on such structures—they "emerge" iff some matter in α_0 "happens" (under the action of the primitive operators or otherwise) to adopt some such special configuration. These are the abstract counterparts of catalysts (particularly enzymes) in real (bio-)chemical systems.

In the study of real chemical systems it is of interest to seek an explanation of the properties and characteristics of catalysis in terms of more fundamental (atomic) interactions. However, for the particular uses we wish to make of the α_0 dynamics, such a more fundamental analysis would be superfluous, and is not attempted. Instead, the properties of emergent operators are simply imposed by

fiat.[6] I may note in passing that the notions of matter conservation and coherent movement (of strongly bonded material segments) make α_0 somewhat reminiscent of von Neumann's (1966a) *kinematic* A-system—though α_0 is, of course, very much simpler.

Holland takes "self-replication" as diagnostic of "life"; the dynamics of α_0 are such that certain complexes (should they arise) may exhibit primitive (but still loosely *genetic*) self-reproducing behaviours.

5.2.2 A Little Formality

In what follows, I shall freely use relevant terminology and notation from the formal theory of computation, as presented, for example, by Lewis & Papadimitriou (1981, especially Section 1.8).

Atoms in α_0 are *formally* defined as *symbols*; the closed loop of atoms is defined as a *string* of exactly R such atomic symbols, which will be referred to as the *state string*; segments and structures are also *strings* (of length less than R) over this same atomic symbol alphabet, normally occurring as *substrings* of the state string;[7] and the operators are *production rules* specifying particular transformations of the state string.

In more detail, the alphabet of atomic symbols is defined as $Z = X \times Y$ (i.e. Z is the cartesian product of two "simple" alphabets X and Y), where:

$$\begin{aligned} X &= \{\mathtt{0}, \mathtt{1}, \mathtt{:}, N_0, N_1, -\} \\ Y &= \{\mathtt{s}, \mathtt{w}\} \end{aligned}$$

$$\begin{aligned} \Rightarrow Z = \{ \quad & (\mathtt{0},\mathtt{s}), (\mathtt{1},\mathtt{s}), (\mathtt{:},\mathtt{s}), (N_0,\mathtt{s}), (N_1,\mathtt{s}), (-,\mathtt{s}), \\ & (\mathtt{0},\mathtt{w}), (\mathtt{1},\mathtt{w}), (\mathtt{:},\mathtt{w}), (N_0,\mathtt{w}), (N_1,\mathtt{w}), (-,\mathtt{w}) \quad \} \end{aligned}$$

[6] This is, in itself, an unusual and interesting (metaphysical) position. The α_0 dynamics might be said to be *irreducible*, to the extent that the properties and behaviours of structures in α_0 are not reducible to properties or behaviours of their "constituent" atoms. However, it should be added that this is still a very weak form of irreducibility, compared to, say, Rosen's (1985b) "complex" systems or Popper's Worlds 1, 2 and 3 (e.g. Popper & Eccles 1977).

[7] Strictly, a segment or structure will not satisfy the technical definition of a "substring" if it spans the atom which is (arbitrarily) designated as the initial atom of the state string. It should be clear that this can be overcome by a minor adjustment to the formal definition of "substring" and this technicality will not, therefore, be discussed further.

We see that each atom (i.e. each atomic symbol) is actually an *ordered pair* of simple symbols, the first taken from the X-alphabet (denoting the element) and the second from the Y-alphabet (denoting the bond state). The state string is then a string of these atoms where each atom is (implicitly) bonded to the next atom to the right. The state string will, of course, be exactly R atoms (ordered pairs) in length.

For many purposes in discussing the α_0 dynamics it will be necessary to refer to just the elements (X-symbols) or the bond states (Y-symbols) in a segment. Two functions are introduced to facilitate this. The first, denoted $\chi()$, extracts the X-symbols from a segment; the second, denoted $\varphi()$, extracts the Y-symbols from a segment. that is, reading the segment (Z-string) from left to right, $\chi()$ maps each Z-symbol onto its X-component:

$$(x, y) \mapsto x$$

and, similarly, $\varphi()$ maps each Z-symbol onto its Y-component:

$$(x, y) \mapsto y$$

5.2.2.1 The Elements

Certain of the elements have similar or related characteristics with respect to the α_0 dynamics. It will therefore prove convenient to group the element (X) symbols into several partially overlapping families or (sub-)alphabets as follows:

$$\begin{aligned} N &\equiv \{N_0, N_1\} \\ A &\equiv \{0, 1, :\} \\ B &\equiv \{0, 1\} \\ M &\equiv N \cup A \\ D &\equiv M - \{:\} \end{aligned}$$

"-" identifies the *null* element, previously mentioned, and is not a member of any of these sub-alphabets.

The N-alphabet serves primarily in the construction of more or less static data storage structures (similar in concept to the A-tapes of the previous chapter); the name N indicates a crude analogy to the function of *nucleotides* in molecular

biology. The A-alphabet serves primarily in the realisation of active structures (emergent operators); the name A indicates a crude analogy to the function of *amino acids.* The B-alphabet is a subset of the A-alphabet; this alphabet is used within emergent operators to code for the operator type and arguments. The name B is a mnemonic for *binary.* The N-alphabet is also, of course, a form of binary alphabet: we shall see that these two distinct binary alphabets are closely related, and this fact partially motivated the particular choice of symbols to represent them. The M-alphabet serves simply to group all the material elements (i.e. as a more concise name for $N \cup A$); the name M is a mnemonic for *material.* Finally, the D-alphabet groups the material elements *other* than the colon element (":"); the name D has no mnemonic significance whatever.

Henceforth I shall refer to atoms whose X-symbol is from the A-alphabet as A-atoms, those whose X-symbol is from the N-alphabet as N-atoms etc. Similarly, a segment consisting exclusively of A-atoms will be called an A-segment etc.

The densities of the separate elements (number of atoms of that element divided by the total number of atoms, R) are parameters of α_0, denoted $\rho(\mathtt{0}), \rho(\mathtt{1})$ etc. The total density of the material elements (M-atoms) is denoted simply by ρ; we must therefore have $\rho(-) = (1 - \rho)$. Typical numerical values, suggested by Holland, which will be used in the empirical investigation are as follows:

$$\rho = \rho(-) = 0.5$$
$$\rho(\mathtt{0}) = \rho(\mathtt{1}) = \rho(\mathtt{:}) = \rho(N_0) = \rho(N_1) = 0.1$$

5.2.2.2 The Bond States

Every atom in α_0 is bonded to the next atom to the right. The state of the bond is denoted by the Y-symbol of each atom: "$\mathtt{s}$" denoting a *strong* bond and "$\mathtt{w}$" denoting a *weak* bond. Note that a strong bond cannot connect with a null atom; this fact constrains the state string in two distinct ways. Firstly, a null atom cannot originate a strong bond, which is to say that the atom $(-, \mathtt{s}) \in Z$ cannot, in fact, arise in α_0. Thus, null atoms will actually occur in α_0 *only* in the form $(-, \mathtt{w})$; we now give this distinguished atomic symbol the special name z_-. Secondly, a null atom cannot terminate a strong bond and thus no segment can arise in α_0 consisting of an atom with a strong bond immediately followed by a null atom (i.e. a segment of the form zz_- where $\varphi(z) = \mathtt{s}$).

5.2.3 The Primitive Operators

There are two primitive operators: Bond Modification (BM) and Exchange (EX). BM is the abstract counterpart of activation, and EX is the abstract counterpart of diffusion.

5.2.3.1 Bond Modification (BM)

BM was originally defined by Holland as follows: on each time step, every bond state in α_0 is (stochastically) updated: each strong bond decays (becoming weak), with probability r; each weak bond becomes strong with probability λr. r and λ are parameters of α_0.

However, as it stands, this is not consistent with the proviso, already stated, that a strong bond cannot connect with a null atom. This does not affect the decay aspect, from strong to weak; but the transformation of weak bonds to strong must strictly be qualified as applying *only* to bonds connecting material atoms (all other bonds, namely the weak bonds connecting null atoms to each other or to material atoms, are thus unaffected by BM).

More formally, BM is defined in terms of two stochastic transformations, or production rules, affecting the state string. The first is the decay from strong to weak:

$$(x, \mathtt{s}) \mapsto (x, \mathtt{w}),\ x \in X$$

This is applied, with probability r, to every atom matching the left hand side (i.e. every atom having a strong bond). The second is the transformation from weak to strong:

$$(x_a, \mathtt{w})(x_b, y) \mapsto (x_a, \mathtt{s})(x_b, y),\ x_a, x_b \in M,\ y \in Y$$

This is applied, with probability λr, to every segment matching the left hand side. r represents bond "stability" (i.e. the probability of a bond decaying from strong to weak). Thus, once a strong bond forms, its lifetime is simply a geometric random variable with parameter r, and the *expected lifetime* of a strong bond is just $1/r$ (neglecting the possible effects of operators other than BM). The typical numerical value used is $r = 10^{-4}$ giving an expected lifetime of a strong bond

of 10^4 time steps. There is no directly analogous result for *weak* bonds because a given weak bond would most likely not be eligible for modification to strong (because it connects to a null atom, transiently or otherwise) throughout its lifetime—and therefore its lifetime could *not* be modeled by any simple geometric random variable.

λ determines (roughly) the "equilibrium ratio" (fixed point of the Markov process implied by **BM**) between weak bonds and strong bonds—*provided* this is interpreted as referring only to bonds connecting material atoms (i.e. which *are* eligible to be strong). Holland argues (in his Theorem 1[8]) that the distribution of structures generated by primitive operators, in isolation, will be *unbiased* (in a precise sense, defined by Holland) iff λ is specified as follows:

$$\lambda \simeq \frac{5(1-\rho)}{3\rho^2} - 1$$

The typical numerical value for λ therefore follows from the value already specified for ρ (0.5), yielding $\lambda = 7/3$. Thus, among all bonds eligible to be strong, approximately two thirds will be expected to be strong, and one third weak, at any given time (at least to the extent that this ratio is determined by **BM**).

In any case, note that the formation and decay of strong bonds are only stochastically related (as represented by λ). The number of strong (or weak) bonds is *not* constant in α_0. To give an extreme example, there is a very small (but still non-zero) probability that, even within a single step, *all* bonds could become weak; or all eligible bonds could become strong, for that matter.

5.2.3.2 Exchange (EX)

The function of **EX** is to provide for a randomised motion or relative rearrangement of the atoms in α_0, with the proviso that any segment consisting exclusively of (necessarily material) atoms which are strongly bonded together will move as a unit.

[8] I should warn that this theorem relies, in turn, on Holland's Lemma 2, and that there are grounds for thinking that the latter is mistaken, both in its result and its derivation—see also section 5.5.2 below. However, it will ultimately be clear that nothing critical relies on this, so I shall accept Holland's analysis at face value just here.

In brief, the idea is that, on each time step, some pairs of adjacent segments, having weak external bonds, exchange positions. The internal bonds of a segment being exchanged may be weak or strong. No bond *states* (internal or external) are altered by the EX operator; what *is* altered is the connectivity between atoms.

However, the details of the EX operator are somewhat complex, as follows.

On each time step, each atom with a weak bond serves, with probability m_1, as the *pivot* for an *exchange operation.* An exchange operation consists of two steps. Firstly, two other atoms with weak bonds are identified as the left and right *limits* of the exchange. This is done by considering first the next atom with a weak bond, to the right of the pivot; this is selected as the right limit, with probability m_2; if it is not selected, then the next atom with a weak bond to the right again is similarly considered, and so on until a limit is determined. The left limit is then established by counting the *same* number of atoms with weak bonds, to the left of the pivot.[9]

In this way, two disjoint, contiguous segments are identified: the left segment consists of all atoms between the left limit and the pivot (excluding the limit, but including the pivot); the right segment consists of all atoms between the pivot and the right limit (excluding the pivot, but including the limit); From the definition, the external bonds of these segments—being those of the left limit, the pivot, and the right limit— are necessarily weak.

If the entire universe is scanned without establishing valid limits (i.e. without identifying two disjoint segments to be exchanged) the exchange operation is aborted—but this should be extremely rare with the typical parameter values.

Conversely, in the normal case, valid left and right segments *are* identified, and these are then swapped (exchanged) with each other, preserving the left to right ordering (and bond states) within the segments. Note that the left and right segments contain the same number of weak bonds but are not, in general, of equal length.

Informally, we may think of an exchange operation as being implemented by "cutting" the right segment out of α_0, and then "splicing" it back into α_0

[9]More concisely: a geometric random variable, with parameter m_2, is sampled; the right and left limits are then established by counting that number of weak bonds to the right and left of the pivot respectively.

immediately to the left of the left segment. These cutting, exchanging, and splicing operations must be pictured as taking place in some space of higher dimensionality, in which α_0 is embedded.

Formally, of course, an exchange operation is a string transformation of the form:

$$z_a z_b z_c z_d z_e \mapsto z_a z_d z_e z_b z_c$$

where:

$$\begin{aligned} z_a, z_c, z_e &\in Z \\ \varphi(z_a), \varphi(z_c), \varphi(z_e) &= \mathtt{w} \\ z_b, z_d &\in Z^* \end{aligned}$$

z_a, z_c, z_e denote, respectively, the left limit, the pivot, and the right limit; $z_b z_c$ is then the left segment, and $z_d z_e$ the right segment, of the exchange operation.

m_1 and m_2 are parameters of α_0. m_1 is roughly analogous to "mean velocity" or "temperature" in chemical systems; the typical value used is 10^{-2}, which is to say that any given atom will serve as a pivot for EX about once in every 100 time steps on which it has a weak bond. m_2 is roughly analogous to "mean free path". Holland does not specify a typical or unique value for m_2 as he suggests that his results are relatively insensitive to it. I shall discuss this in more detail in section 5.5 below.

Note that the EX operator does *not* emulate anything even approximating to Newtonian mechanics. Force, velocity, momentum or kinetic energy are not even meaningful concepts in α_0. In particular, there is no notion of conservation of (kinetic) energy.

Since exchange operations with distinct pivots might interfere with each other (if their limits overlapped) Holland stipulated that the various exchanges should occur sequentially from the "leftmost" pivot to the "rightmost".

This stipulation implicitly requires that α_0 be finite: otherwise a single time step of this (strictly sequential) EX operator could never be completed; but this is not a substantive issue since any practical realisation would have to be finite anyway. More significantly, this stipulation also implicitly requires that α_0 be *bounded* (and not, for example, circular as assumed up to now); otherwise "leftmost" and "rightmost" atoms would not be well defined. However, Holland gives

no detailed discussion of the exact behaviour of α_0 at these implied boundaries, neither in the specific context of EX, nor elsewhere.

Furthermore, this mechanism for the ordering of the exchange operations still requires one further clarification. As stated, it is not clear whether, for the purposes of EX, α_0 should be scanned left to right just once (deciding, at each atom with a weak bond, whether to carry out an exchange, and, if so, then immediately carrying it out) or twice (first to decide which atoms should serve as pivots, and then a second time to actually carry out the exchanges). The former approach seems more straightforward, but suffers from a subtle defect: with such an approach, on any single time step, some atoms may not be even *considered* as candidate pivots for an exchange operation, and some others may actually be considered several times. This would arise whenever the outer limits are more than one weak bond away from the pivot: for then some atom(s) with weak bonds, which have not yet been considered as candidate pivots, will be shifted to the left of the current pivot, and will therefore be passed over; and conversely, some atoms with weak bonds to the left of the pivot, which have (presumably) already been considered as possible pivots, will be shifted to the right of the current pivot and will be considered again. Even more convoluted scenarios can, of course, be imagined. This defect is avoided with the alternative approach of scanning the complete space twice, for then the pivots are *all* identified before *any* exchanges are carried out. While Holland did not discuss this issue in such detail, it seems that this must have been his intended ordering mechanism.

In any case, we shall subsequently see that, in the implementation of α_0 to be described, there are good reasons for ultimately adopting a somewhat different (and simpler) approach to the EX operator. This will, of course, retain the required statistical characteristics, and will still involve implementing potentially conflicting exchange operations sequentially; but it will improve the execution speed, and, as an added bonus, it will transpire to be immune to the kind of ordering difficulties just described, so that the issue of scanning from a leftmost boundary to a rightmost boundary (however many times) will not arise. This will sidestep the requirement for special "boundary" behaviour completely, and allow a symmetrical (unbounded) treatment of α_0 as previously assumed. Similar comments apply to certain aspects of the behaviour of the emergent operators.

5.2.4 The Emergent Operators

The "activation" conditions for BM and EX are such that they are bound to operate in arbitrary states of α_0: BM is guaranteed to establish a population of atoms with weak bond states, and this, in turn, guarantees that the conditions for EX to operate will be satisfied. In particular, this means that these operators will be effective even in the most "disordered" or "primitive" states of α_0—and it is for this reason that they are termed *primitive.*

In contrast to this, the remaining operators are contingent on more complicated activation conditions; it seems therefore that they will have a substantive effect on the α_0 dynamics only if such (comparatively) special states should occur.[10] It is for this reason that these other operators are termed *emergent,* in the sense that they (or their effects) will not be manifest in arbitrary states of α_0, but may become manifest (emerge) if pre-existing operators (the primitive operators, in the first instance) should happen to cause the relevant special states to arise.

As already mentioned, the effects of emergent operators are roughly analogous to the functions of catalysts and enzymes in real biochemical systems. Formally, we shall identify or label the emergent operators with certain relatively invariant aspects of their activation conditions—specifically, with certain segments, or classes of segments, which remain essentially unaltered through a cycle of transformations associated with a particular operator. We shall then say that these distinguished segments *are* the emergent operators, or E-OPs.

While, in principle, the set of distinct E-OPs is infinite, they are all more or less similar in operation. This will allow the specification of their behaviour to be streamlined. In particular, the E-OPs will all be classified into just two major groups, each characterised by a certain "typical" cycle of transformations, or "reaction cycle". Loosely speaking, an E-OP can be considered as a finite state automaton, embedded in α_0, which will automatically transit through a certain typical cycle of "states".

However, it is important to emphasise, even at this point, that these "typical" cycles are simply a *device,* adopted to allow a more concise and systematic

[10]The task of quantifying just how special, or otherwise, these conditions are is a major element in any attempted analysis of α_0.

description of the E-OPs: it should not be taken to imply that, *in fact*, E-OPs can only arise or emerge in some "initial" state, or that an E-OP "reaction cycle" will necessarily run to completion. In particular, an E-OP may be spontaneously created, modified, or destroyed in any arbitrary state—i.e. at any arbitrary point in its "cycle"—by the effects of the EX *primitive* operator. This fact is referred to only obliquely by Holland; but it will transpire that it is critical to the overall behaviour of α_0.

5.2.4.1 The *Codon String* Function $\pi()$

Before attempting to characterise the E-OPs proper it is necessary to define another, different, kind of object, termed a *codon structure.*

Informally, a codon structure is the analog in α-Universe of a polynucleotide in molecular biology. It is loosely defined as any structure (null delimited segment) whose leftmost material atom (at least) is an N-atom. More formally, we define the set of codon structures as a language, $L_{CS} \subset Z^*$, as follows:

$$\begin{aligned} L_{CS} \equiv \{ \quad & z_{CS} \in Z^*, \\ & z_{CS} = z_- z_a z_b z_-, \\ & \chi(z_a) \in N^+, \\ & \chi(z_b) \in (M^* - NM^*) \quad \} \end{aligned}$$

(Note that the notation N^+ is a short form for NN^*, which is to say the set of all strings over N having at least one symbol.)

In words: a codon structure is any structure whose material segment has an N-segment prefix (z_a above). This prefix will, necessarily, be uniquely delimited at the right—either by the null atom (z_-) terminating the structure or by an A-atom (i.e. some M-atom which is not an N-atom). In the latter case, this A-atom, and any M-atoms following it, will be referred to as a "garbage suffix" to the codon structure (z_b above).

It is convenient to define the function:

$$\begin{aligned} \pi : L_{CS} \quad &\rightarrow \quad N^+ \\ z_{CS} = z_- z_a z_b z_- \quad &\mapsto \quad \chi(z_a) \end{aligned}$$

where the decomposition $z_{CS} = z_- z_a z_b z_-$ is as introduced in the definition of L_{CS}

above. That is, $\pi(z_{CS})$ yields the N-string corresponding to the N-segment prefix of the codon structure. Such an N-string will be termed a *codon string.*

Examples of codon structures (members of L_{CS}), and their corresponding codon strings, might be the following:

z_{CS}	$\chi(z_{CS})$	$\pi(z_{CS})$
$z\_(N_1,\mathtt{s})(N_0,\mathtt{w})z\_$	$-N_1N_0-$	N_1N_0
$z\_(N_1,\mathtt{w})(0,\mathtt{s})(N_1,\mathtt{s})(N_1,\mathtt{w})z\_$	$N_10N_1N_1$	N_1
$z\_(N_0,\mathtt{w})(:,\mathtt{w})(0,\mathtt{w})(0,\mathtt{w})(:,\mathtt{w})z\_$	N_0:00:	N_0
$z\_(N_0,\mathtt{s})(N_1,\mathtt{s})(N_0,\mathtt{s})(1,\mathtt{s})(N_1,\mathtt{w})z\_$	$N_0N_1N_01N_1$	$N_0N_1N_0$

Codon structures and E-OPs are mutually exclusive—i.e. no codon structure is an E-OP and vice versa. Codon structures are thus more or less "static"—in the sense that the only dynamic behaviour they exhibit *in themselves* is that implied by the primitive operators. We shall see that codon structures (or, at least, the codon strings thereof) play the rôle of "data" objects operated upon by the "programs" represented by the E-OPs. As mentioned earlier, they are somewhat analogous to the "A-tapes" of the previous chapter.

5.2.4.2 The *Binding* Function $\alpha()$

In general, if an E-OP is to operate on some codon structure, it must first identify a "suitable" structure, and then attach this structure (or, at least, the material segment of it) onto itself. I shall refer to this process as *selective binding.* Binding is *selective* in that there will be an embedded B-segment within the E-OP (termed the *argument* of the E-OP) which will constrain the binding—the selected codon structure must "match" this B-segment. This matching is insensitive to bond states either in the E-OP or the codon structure—it is based purely on the X-symbols in both cases. The matching condition is expressed via a mapping from N-strings (specifically, codon strings) to B-strings, called the *binding function,*[11] and denoted by:

$$\alpha : N^* \rightarrow B^*$$

[11] Holland actually describes this function as yielding an "anticode" for a given N-string; I find this confusing, as it seems to imply some relationship with the "coding" function $\gamma()$ (to be described in the next section), whereas there is no such relationship. $\alpha()$ and $\gamma()$ are quite independent both in definition and application. In particular, it is *not* the case that $\alpha = \gamma^{-1}$. Thus I prefer not to propagate the term "anticode" further.

$\alpha()$ is defined as follows: reading the N-string from left to right, each N-symbol is mapped onto a single B-symbol, according to:

$$N_0 \mapsto \texttt{0}$$
$$N_1 \mapsto \texttt{1}$$

Thus, we can say, for example:

$$\alpha(N_0N_1N_0N_1) = \texttt{0101}$$
$$\alpha(N_0N_1N_1N_0N_0) = \texttt{01100}$$

It is, of course, the relationship implied by $\alpha()$ which originally motivated the particular choice of symbols for the N-alphabet. A codon structure $z_{\texttt{CS}} \in L_{\texttt{CS}}$ will then be said to match a relevant B-segment u iff $\alpha(\pi(z_{\texttt{CS}}))$ contains $\chi(u)$ as a prefix. This condition will be denoted $z_{\texttt{CS}} \bowtie u$. More concisely, we require that:

$$z_{\texttt{CS}} \bowtie u \iff \alpha(\pi(z_{\texttt{CS}})) = x_a x_b,\ x_a = \chi(u) \in B^+,\ x_b \in B^*$$

I may note in passing that, since $\alpha()$ is bijective, its inverse, $\alpha^{-1}()$, is well defined; this proves convenient when it comes to practical realisation of selective binding, as it is somewhat easier to implement the matching condition $z_{\texttt{CS}} \bowtie u$ by expressing it, in terms of the segments defined above, as the condition that $\pi(z_{\texttt{CS}})$ must contain $\alpha^{-1}(\chi(u))$ as a prefix.

5.2.4.3 The *Decoding* Function $\gamma^{-1}()$

The definition of certain E-OPs (the *decode* type) will involve "constructing" an A-segment based on interpreting a given codon string as a "description" of it. The general idea is that codon strings may act as (more or less) quiescent descriptions of certain A-segments; these A-segments themselves are, in general, *not* quiescent (i.e. they may be, or become, E-OPs). The mapping from an A-segment to its description (codon string) is referred to in (Holland 1976) as the "coding" function, and is denoted by $\gamma()$. This function specifies a coding *only* for the X-symbols of the atoms in the A-segment—it is insensitive to the bond states, and is thus of the form:

$$\gamma : A^* \rightarrow N^*$$

i.e. a mapping from an A-string to an N-string.

The definition[12] of $\gamma()$ is that, reading the A-string from left to right, each A-symbol maps onto a pair of N-symbols according to:

$$\begin{aligned} \texttt{0} &\mapsto N_0N_0 \\ \texttt{1} &\mapsto N_0N_1 \\ \texttt{:} &\mapsto N_1N_0 \end{aligned}$$

Thus, we can say, for example:

$$\begin{aligned} \gamma(\texttt{010:}) &= N_0N_0N_0N_1N_0N_0N_1N_0 \\ \gamma(\texttt{0:100}) &= N_0N_0N_1N_0N_0N_1N_0N_0N_0N_0 \\ \gamma(\texttt{::::}) &= N_1N_0N_1N_0N_1N_0N_1N_0 \end{aligned}$$

This relationship between A-strings and N-strings—i.e. this particular coding of A-strings into N-strings—"exists" *only* in the sense that certain `E-OP` dynamics reflect it; nonetheless, it is convenient to define it here, separately from the detailed description of the `E-OP`s.

Of course, the function that must be used in constructing an A-segment from its description is not $\gamma()$, but its inverse—i.e. the *decoding* function, $\gamma^{-1}()$. Unfortunately, $\gamma()$, as so far defined (following Holland), is not bijective—there are many N-strings which do not code for any A-string. This arises, for example, if the N-string has an odd length, or if, when it is split into pairs from left to right, the pair N_1N_1 occurs. Thus, the "correct" definition of $\gamma^{-1}()$ is somewhat arbitrary. Holland explicitly specifies that $\gamma^{-1}()$ should map odd length N-strings by ignoring the final N-symbol in the string (thus effectively making it of even length). However, Holland does *not* specify how $\gamma^{-1}()$ should deal with the pair N_1N_1. For the work described here N_1N_1 was (arbitrarily) mapped onto ":"—thus, both N_1N_0 and N_1N_1 effectively code for the same A-symbol. The definition of $\gamma^{-1}()$ is then fully characterised by the following mappings:

$$\begin{aligned} N_0N_0 &\mapsto \texttt{0} \\ N_0N_1 &\mapsto \texttt{1} \\ N_1N_0 &\mapsto \texttt{:} \\ N_1N_1 &\mapsto \texttt{:} \end{aligned}$$

[12]This is not literally the definition given by Holland: the latter evidently incorporated a typographical error, as it showed the *same* coding for both `0` and `1`.

5.2.4.4 Searching for Raw Materials

At various points it will be necessary for an E-OP to search for "raw materials", so that it can effect particular transformations while still respecting the matter conservation "laws" of α_0. This will involve searching for a segment of a particular, specified, form. A standard searching procedure is used, which is somewhat similar to that described for establishing the limits of an exchange operation in section 5.2.3.2 above.

The search proceeds outward from the E-OP (concurrently to both the left and right); when a suitable segment is located (on the left *or* right), then, with probability m_2, this segment is selected and the search terminates; otherwise the search continues outward. Effectively a geometric random variable, with parameter m_2, is sampled, yielding a count—say k; the k'th nearest suitable segment is then selected (if it exists).

This procedure is completed within a single time step, regardless of how far the search has to proceed—up to, and including, searching the complete space. I may note that this ability to search arbitrarily far in a constant (α_0) time is a particularly counter-intuitive feature of α_0. Holland does suggest that consideration should ultimately be given to more "realistic" procedures (Holland 1976, Section 5), though I shall not pursue that here.

This search procedure may *fail*; that is, the search may exhaust α_0 without having selected any segment. This would certainly occur if *no* suitable segment existed in the (finite) α_0, and might occur even if one or more suitable segments exist but they are all passed over (which can happen with probability $1 - m_2$ for each such segment).

In any case it is stipulated that if a search fails, for whatever reason, the E-OP in question will have no effect on α_0 for that time step; this will typically mean that the search procedure will then be repeated, afresh, on the next time step.[13]

[13]Holland states this point *explicitly* only for one special case; but I have applied the same principle, *mutatis mutandis*, to all comparable cases.

5.2.4.5 Outline E-OP Syntax

All E-OPs supported in α_0 are structures and are, more particularly, members of the language $L_{\texttt{E-OP}}$, formally defined as follows:

$$\begin{aligned} L_{\texttt{E-OP}} \equiv \{ \quad & z_{\texttt{E-OP}} \in Z^*, \\ & z_{\texttt{E-OP}} = z\_uvz\_, \\ & u = u_a u_b u_c, \\ & \chi(u_a) \in B, \\ & \chi(u_b) = :, \\ & \chi(u_c) \in B^+, \\ & \chi(v) \in (M^* - BM^*) \quad \} \end{aligned}$$

Less formally, an E-OP is a structure (a null delimited segment), whose material segment can be (uniquely) decomposed into two distinct parts, u and v. We shall see that u remains essentially invariant through the cycle of transformations associated with the E-OP, and represents its fixed "program" part. u can be further decomposed into a B-atom u_a, called the operator *type*, a colon atom u_b which is just a separator,[14] and a B-segment u_c, called the operator *argument*. u_c is stipulated not to be empty ($e \notin B^+$) but is otherwise an arbitrary B-segment.[15]

The material segment v is called the *operand*; it is progressively modified through the cycle of transformations and may be regarded (roughly) as the "data" being processed by the E-OP, or as the record of the instantaneous "state" of the E-OP. At this point v is permitted to be an arbitrary material segment ($\chi(v) \in M^*$) *except* that its leftmost atom (if any) must not be a B-atom ($\chi(v) \notin BM^*$). In this way we are guaranteed that, whether v is empty or not, the argument u_c (and thus u itself) will be uniquely delimited on the right, being immediately followed by an atom whose X-symbol is *not* in B (namely the first atom in v, if it exists, or, if v is empty, the right null atom delimiter of the structure). This ensures that the decomposition of the E-OP as $z\_uvz\_$ is unambiguous.

[14] In fact, this separator is not strictly required in α_0, because the length of u_a is fixed; however, Holland intended this syntax to be extensible to more complicated α-Universes, where u_a would be of variable length ($u_a \in B^+$), hence the inclusion of u_b even in α_0.

[15] There are conceivable alternatives to requiring simply that the argument be non-empty: we might permit an empty argument, or we might impose a minimum length greater than one. Holland is not entirely clear on this issue: he explicitly refers to the possible use of an argument length of two, so the minimum should not be more that this—but it could still be zero, one or two without contradicting anything in Holland's original paper. I simply note the convention I have adopted—i.e. that the minimum valid argument length is one.

In practice, the definition of the E-OP dynamics is such that only a small number of really distinct kinds of transformation can be effected by a given E-OP. Thus, even though the set of possible operands v is very large, these will be classified into a small number of distinct classes for the purposes of defining the resulting transformations. This classes will, of course, be disjoint. Operands which do not fall into *any* of these classes are considered, by default, to be representatives of a "halt" class—which is to say that an E-OP, in such a state, will not cause any transformations at all.

Examples of members of $L_{\texttt{E-OP}}$ would be segments having the following images under $\chi()$:

$$-0{:}1-$$
$$-1{:}0N_0N_1{::}-$$
$$-0{:}0000010N_0{:}N_1N_110{:::}-$$
$$-1{:}100{:}100-$$

E-OPs are classified into types on the basis of the X-symbol of the B-atom u_a in the definition above; since $B = \{0, 1\}$, this yields just two distinct types. An E-OP with $\chi(u_a) = 0$ will be called a *copy* type (henceforth denoted CP), and an E-OP with $\chi(u_a) = 1$ will be called a *decode* type (henceforth DC). The behaviour of E-OPs of these two types will be defined in detail in the following sections.

5.2.4.6 The CP E-OPs

In brief, the CP reaction cycle consists of locating and incorporating a codon structure which matches the argument; copying the N-segment prefix part of this (effectively the codon string proper), by locating and incorporating free N-atoms in the appropriate order; and finally, dividing (incorporating null atoms) in such a way as to reconstitute the original codon structure and delimit the copy thereof; this also returns the E-OP to the initial state in the cycle. In principle, a given E-OP may repeat this cycle indefinitely although, of course, it will sooner or later be broken up through the action of the other operators (especially BM and EX).

Loosely then, the essential behaviour of a CP E-OP can be best understood by examining its image under $\chi()$, through a complete reaction cycle. A "typical"

sequence of this sort would be as follows:

t $\quad -$0:010$-$

$t+1$ $\quad -$0:010:$-$

$t+2$ $\quad -$0:010:$N_0N_1N_0N_0-$

$t+3$ $\quad -$0:010:$N_0N_1N_0N_0$:$-$

$t+4$ $\quad -$0:010N_0:$N_1N_0N_0$:N_0-

$t+5$ $\quad -$0:010N_0N_1:N_0N_0:N_0N_1-

$t+6$ $\quad -$0:010$N_0N_1N_0$:N_0:$N_0N_1N_0-$

$t+7$ $\quad -$0:010$N_0N_1N_0N_0$::$N_0N_1N_0N_0-$

$t+8$ $\quad -$0:010$-N_0N_1N_0N_0-$:$-$:$-N_0N_1N_0N_0-$

Note that, in general, a CP E-OP will require at least $\ell + 4$ time steps to complete its reaction cycle, where ℓ is the length of the N-segment prefix in the codon structure which is copied ($\ell = 4$ in the example above). The time required may, of course, be longer than this—for example if the search for relevant raw materials should fail at any point.

This outline of the "normal" reaction cycle of a CP E-OP is, in fact, *all* that Holland specified of its behavior. As already noted, however, if one is interested in building a practical realisation of α_0, it is necessary to consider not just this kind of normal reaction cycle, but also all the other possible E-OP states which might conceivably arise. The rest of the discussion of the CP E-OPs below is therefore concerned with giving a *complete* and *formal* definition of the transformations they effect, such that reaction cycles of the kind loosely implied above will, in fact, result.

From the discussion already given, all CP E-OPs must be members of the language $L_{\mathtt{CP}} \subset L_{\mathtt{E-OP}}$ defined by:

$$
\begin{aligned}
L_{\mathtt{CP}} \equiv \{ \quad & z_{\mathtt{CP}} \in Z^*, \\
& z_{\mathtt{CP}} = z\_uvz\_, \\
& u = u_a u_b u_c, \\
& \chi(u_a) = 0, \\
& \chi(u_b) = :, \\
& \chi(u_c) \in B^+, \\
& \chi(v) \in (M^* - BM^*) \quad \}
\end{aligned}
$$

The particular transformations implemented by a CP E-OP are determined by the instantaneous state—i.e. by the operand segment v. The transformations associated with any given operand are completed in one time step. This will result in a new operand; this will "typically" either be a member of the next operand class in the reaction cycle, or will still be a member of the same operand class. However, I emphasise again that an E-OP can, in principle, be "initially" formed with an operand of arbitrary class, and the relevant transformations will, of course, then proceed from there; and equally, the operand may fail to be transformed from one class to to the next in the reaction cycle for various reasons, despite the cycle being described as "typical" or "normal".

5.2.4.6.1 CP Operand Class 0

There is exactly one class 0 operand, namely the empty string:

$$v = e$$

Informally, the class 0 transformations involve locating a free colon atom and splicing this into the E-OP, strongly bonding it to the right end of the structure (i.e. between the rightmost existing M-atom, and the right null atom delimiter). This colon atom will be subsequently used to mark the current position in an N-segment (derived from a codon structure) while it is being copied; it will be referred to below as the *position marker.*

The "location" of the free colon atom is an example of a search for "raw materials", which has already been detailed in section 5.2.4.4. As noted there, this procedure may, in general *fail*; and, should that happen, then *none* of the transformations described here will be effected. In particular, the operand itself will remain unchanged, and typically, therefore, the procedure will be attempted afresh on the next time step (etc.). These considerations will apply in every case where reference is made to locating particular kinds of segment, and will not be repeated further.

More formally, let the decomposition of the segment u of the E-OP, into its constituent atoms, be denoted in the normal way by:

$$u = u(1)u(2)\ldots\ u(|u|),\ u(i) \in Z$$

Note that $|u|$ denotes the length of u. $u(|u|)$ will thus denotes the rightmost atom of u. It is necessarily the case that $\varphi(u(|u|)) = \mathtt{w}$ (i.e. this atom has a weak bond state) because the next atom to the right is null ($= z\_$).

Let z denote a free colon atom:

$$z = z\_(:, \mathtt{w})z\_$$

The following transformations are then triggered by a class 0 operand:

$$\begin{aligned} z &\mapsto z\_z\_ \\ v &\mapsto (:, \mathtt{w}) \\ \varphi(u(|u|)) &\mapsto \mathtt{s} \end{aligned}$$

5.2.4.6.2 CP Operand Class 1

Again, there is exactly one class 1 operand, namely a single colon atom (necessarily with a weak bond):

$$v = (:, \mathtt{w})$$

Informally, the class 1 transformations involve locating a codon structure matching the argument of the E-OP, and splicing the material segment of this into the E-OP, strongly bonding it to the right of the colon atom v. The internal bonding of the material segment extracted from the codon structure is not altered.

More formally, let $z_{\mathtt{CS}}$ be a codon structure:

$$\begin{aligned} z_{\mathtt{CS}} = \; & z\_z_a z_b z\_, \\ & \chi(z_a) \in N^+, \\ & \chi(z_b) \in (M^* - NM^*) \end{aligned}$$

We require that $z_{\mathtt{CS}}$ match the E-OP argument, in the sense of the $\alpha()$ function as described in section 5.2.4.2; that is:

$$z_{\mathtt{CS}} \bowtie u_c$$

The following transformations are then triggered by a class 1 operand:

$$\begin{aligned} z_{\mathtt{CS}} &\mapsto z\_z\_ \\ v &\mapsto (:, \mathtt{s})z_a z_b \end{aligned}$$

5.2.4.6.3 CP Operand Class 2

Class 2 operands are defined by the condition:

$$\begin{aligned} v = \; & v_a v_b v_c v_d, \\ & \chi(v_a) \in (D^* - BD^*) \\ & \chi(v_b) = :, \\ & \chi(v_c) \in ND^*, \\ & v_d = e \end{aligned}$$

(The reason for including v_d here will become clear subsequently.)

It should be clear that this decomposition, and similar ones which follow in subsequent sections, will be unique. This point will not, therefore, be repeated.

In words, the operand is of class 2 if it has exactly one colon atom in it, and the atom immediately to the right of this is an N-atom.

Note carefully that the segment denoted v_c must not contain any colon atoms (recall that D is the set of material elements *exclusive* of the colon element). Now, there is no constraint in the definition of codon structures which guarantees that their material segments (which is what v_c is typically derived from) will not contain colon atom(s) in the garbage suffix. Thus, the "typical" transition from a class 1 operand to a class 2 operand might be subverted. It would, arguably, be preferable to define the class 1 transformations to reduce or eliminate this possibility, by stipulating that *only* the N-segment prefix of a codon structure should be incorporated into the E-OP, rather than the complete material segment (i.e. the garbage suffix, if any, would be discarded by the action of the class 1 operand). This would also, incidentally, slightly simplify the definitions of later operand classes and their transformations. However, while Holland is not completely unambiguous on this point, there is a strong implication in his treatment that the garbage suffix should not be discarded in this way, so I leave the class 1 transformations as they have already been stated.

Informally, the class 2 transformations are similar to those for the class 0 operand: a free colon atom is located and spliced into the E-OP, strongly bonding it to the right of the rightmost existing M-atom. This colon atom will be subsequently used to separate the original N-segment from the copy.

More formally, let z denote a free colon atom:

$$z = z\_(:, \mathtt{w})z\_$$

The following transformations are then triggered by a class 2 operand:

$$\begin{aligned} z &\mapsto z\_z\_ \\ v_d &\mapsto (:, \mathtt{w}) \\ \varphi(v_c(|v_c|)) &\mapsto \mathtt{s} \end{aligned}$$

(The reason for including $v_d = e$ in the prior decomposition of v now becomes apparent: it allows the second of these transformations to be expressed relatively concisely.)

5.2.4.6.4 CP Operand Class 3

Class 3 operands are defined by the condition:

$$\begin{aligned} v = \; & v_a v_b v_c v_d v_e v_f v_g, \\ & \chi(v_a) \in (D^* - BD^*), \\ & \chi(v_b) = :, \\ & \chi(v_c) \in N, \\ & \chi(v_d) \in D^*, \\ & \chi(v_e) = :, \\ & \chi(v_f) \in M^*, \\ & \chi(v_g) = e \end{aligned}$$

In words, the operand is of class 3 if it contains at least two colon atoms, and the atom immediately to the right of the first colon atom (the position marker) is an N-atom.

Informally, the class 3 transformations involve "copying" a single N-atom; that is, a free N-atom is located, matching the N-atom immediately to the right of the position marker, and this is spliced into the E-OP, strongly bonding it to the right of the rightmost existing M-atom. The position marker is also exchanged with the N-atom which has just been copied. In general, this will produce an operand which is still of class 3 unless and until the position marker is moved to a point where the next atom is not an N-atom (typically, it will either be the

colon atom added by the class 2 operand, or the first atom of the garbage suffix of the original codon structure). In this way, the E-OP should continue to have a class 3 operand, and will continue copying one N-atom per time step, until the embedded codon string is completely copied.

More formally, let z denote the required free N-atom:

$$z = z\_(\chi(v_c), \mathtt{w})z\_$$

The following transformations are then triggered by a class 3 operand:

$$\begin{aligned} z &\mapsto z\_z\_ \\ v_g &\mapsto (\chi(v_c), \mathtt{w}) \\ \varphi(v_f(|v_f|)) &\mapsto \mathtt{s} \\ v_c &\mapsto v_b \\ v_b &\mapsto v_c \end{aligned}$$

5.2.4.6.5 CP Operand Class 4

Class 4 operands are defined by the condition:

$$\begin{aligned} v = \; & v_a v_b v_c v_d v_e, \\ & \chi(v_a) \in (D^* - BD^*), \\ & \chi(v_b) = :, \\ & \chi(v_c) \in (D^* - ND^*), \\ & \chi(v_d) = :, \\ & \chi(v_e) \in M^* \end{aligned}$$

In words, the operand is of class 4 if it contains at least two colon atoms, and the atom immediately to the right of the first colon atom (the position marker) is *not* an N-atom.

Informally, the class 4 transformations involve breaking up and ejecting the operand, typically establishing two distinct, but identical (in the sense of $\pi()$), codon structures (and possibly an additional garbage structure), separated from the E-OP. One of the codon structures is effectively the "original", and the other is the newly constructed "copy". By ejecting the existing operand, the transformed operand (namely the empty string e) will be of class 0 again, and the reaction cycle has been closed.

These transformations are achieved by locating "free" null atoms and splicing them into the (former) E-OP in appropriate locations. Bond states will be forced to weak wherever necessary. The precise details vary depending on whether certain substrings which, by the definition above, are allowed to be empty, are, in fact, empty (since clearly, in such a case, there is no point in "delimiting" such empty strings by null atoms).

The transformations associated with the class 4 operands are the most complicated of all, and involve at least two, and perhaps as many as five, distinct sets of related transformations. These sets must be implemented sequentially, and each separately involves a "location" procedure, which may potentially fail. It is stipulated however, that *all* required location procedures must be successfully completed before *any* transformations are carried out; and if any location procedure actually fails then *no* transformations will be carried out.

It is convenient to regard each distinct set of transformations as a case of a generic "null-insertion" set of transformations which inserts a single null atom immediately to the right of a specified segment, say z. Let $z = z_a z_b$ where $z_b = e$. A segment $z_c = z\_z\_$ is first located. The null-insertion transformations, applied to z, are then defined as follows:

$$\begin{aligned} z_c &\mapsto z\_ \\ z_b &\mapsto z\_ \\ \varphi(z_a(|z_a|)) &\mapsto \mathtt{w} \end{aligned}$$

The complete set of transformations triggered by a class 4 operand then consists of the sequential application of the null-insertion transformations to the segments u, v_a (iff $v_a \neq e$), v_b, v_c (iff $v_c \neq e$), and v_d (iff $v_e \neq e$), in that order.

5.2.4.7 The DC E-OPs

The behaviours of the DC E-OPs are extremely similar to those of the CP E-OPs. The discussion here will therefore concentrate just on those aspects in which the two E-OP types *differ*.

The DC reaction cycle consists of locating and incorporating a codon structure which matches the argument; decoding (in the sense of $\gamma^{-1}()$) the N-segment prefix part of this (effectively the codon string proper), by locating and incorporating free A-atoms in the appropriate order; and finally, dividing (incorporating null atoms) in such a way as to reconstitute the original codon structure and delimit the decoded version thereof; this also returns the E-OP to the initial state in the cycle. In principle, a given E-OP may repeat this cycle indefinitely although, of course, it will sooner or later be broken up through the action of the other operators (especially BM and EX).

As for CP, the essential behaviour of a DC E-OP can be best understood by examining its image under $\chi()$, through a complete reaction cycle. A "typical" sequence of this sort would be as follows:

$$
\begin{array}{ll}
t & -0{:}010- \\
t+1 & -0{:}010{:}- \\
t+2 & -0{:}010{:}N_0N_1N_0N_0- \\
t+3 & -0{:}010{:}N_0N_1N_0N_0{:}- \\
t+4 & -0{:}010N_0N_1{:}N_0N_0{:}1- \\
t+5 & -0{:}010N_0N_1N_0N_0{::}10- \\
t+6 & -0{:}010-N_0N_1N_0N_0-{:}-{:}-10-
\end{array}
$$

Thus, in general, a DC E-OP will require at least $\ell/2+4$ time steps to complete its reaction cycle, where ℓ is the length of the N-segment prefix in the codon structure which is decoded ($\ell = 4$ in the example above). Again, the time actually required may be longer than this.

All DC E-OPs must be members of the language $L_{\mathtt{DC}} \subset L_{\mathtt{E-OP}}$ defined by:

$$
\begin{aligned}
L_{\mathtt{DC}} \equiv \{ \quad & z_{\mathtt{DC}} \in Z^*, \\
& z_{\mathtt{DC}} = z\_uvz\_, \\
& u = u_a u_b u_c, \\
& \chi(u_a) = 1, \\
& \chi(u_b) = :, \\
& \chi(u_c) \in B^+, \\
& \chi(v) \in (M^* - BM^*) \quad \}
\end{aligned}
$$

As with CP, the operands of the DC E-OP are grouped into classes which effect essentially similar transformations. The operand classes 0, 1 and 2, and the resulting transformations, are identical for both CP and DC and will not be repeated. The definitions of the remaining 2 classes, and their transformations, differ slightly, and will be discussed individually.

5.2.4.7.1 DC Operand Class 3

The definition of this class is identical to CP class 3, *except* that the codon string part must consist of at least two atoms, where one sufficed in the CP case. In terms of the decomposition defined for CP, we require that:

$$\chi(v_c) \in N^2$$

The resulting transformations are somewhat similar to those of CP, but with several significant differences. A free A-atom (rather than N-atom) is located and spliced into the E-OP. The X-symbol of this A-atom must equal the decoded version of the next pair of N-atoms in the codon string part of the E-OP (in the sense of the $\gamma^{-1}()$ function). Finally, the position marker must be moved *two* positions to the right rather than just one. As with CP, the new operand will again typically be of (DC) class 3, and transformations of this sort will be iterated until the codon string part of the E-OP is exhausted (has a length less than two in this case).

More formally, let z denote the required free A-atom:

$$z = z\_(\gamma^{-1}(\chi(v_c)), \mathtt{w})z\_$$

The following transformations are then triggered by a DC class 3 operand:

$$\begin{aligned} z &\mapsto z\_z\_ \\ v_g &\mapsto (\gamma^{-1}(\chi(v_c)), \mathtt{w}) \\ \varphi(v_f(|v_f|)) &\mapsto \mathtt{s} \\ v_c &\mapsto v_b \\ v_b &\mapsto v_c \end{aligned}$$

5.2.4.7.2 DC Operand Class 4

The definition of this class is identical to CP class 4, except that the garbage remaining from the codon string is now allowed to include a single initial N-atom. In terms of the decomposition defined for CP class 4, we require that:

$$\chi(v_c) \in (D^* - N^2D^*)$$

With this modification of the definition of v_c, the resulting transformations are then identical to those defined for CP Class 4 operands.

5.3 "Life" in α_0?

Consider a complex in α_0 comprising 8 structures having the following images under $\chi()$:

$$\begin{array}{ll} -\texttt{0:001}- & -N_0N_0N_1N_0N_0N_0N_0N_0N_0N_1- \\ -\texttt{0:011}- & -N_0N_0N_1N_0N_0N_0N_0N_1N_0N_1- \\ -\texttt{1:001}- & -N_0N_1N_1N_0N_0N_0N_0N_0N_0N_1- \\ -\texttt{1:011}- & -N_0N_1N_1N_0N_0N_0N_0N_1N_0N_1- \end{array}$$

For reasons which will become more clear subsequently, this complex will be referred to as FullSR.

It will be observed that the structures represented in the left column are all E-OPs, the first two of type CP, the second two of type DC, while the structures represented in the right column are all codon structures. The first CP can bind with the first two codon structures and is thus capable of copying them; the second CP can bind to, and thus copy, the remaining two codon structures. Similarly, the first DC can bind to, and decode, the first two codon structures, and the second DC can bind to, and decode, the remaining two codon structures. Finally, it can be easily verified that, when the four codon structures are decoded they actually yield precisely the four E-OPs represented in the left column.

Prima facie, then, the complex FullSR is capable of a form of "self-reproduction": it is an example, in α_0, of a roughly von Neumann style, genetically based, *A-reproducer*.

Admittedly, complexes in α_0 lack some of the coherence or unity of the A-machines considered in the previous chapter. That is, given a population of

structures there is a certain arbitrariness in identifying which of these constitute distinct complexes. When we say that `FullSR` is self-reproducing, what we mean is that, given a single instance of it, this should result in the generation of a large population of structures which contains many instances of each of the component structures of `FullSR`; but it is then quite arbitrary which of these structures should be grouped together as "instances" of `FullSR` itself. There is actually a rather fundamental issue at stake here: to anticipate somewhat, we can roughly regard the recursive interaction of the structures making up `FullSR` as realising a form of genetic self-reproduction *or* as realising a quasi-autopoietic organisation—*but not as doing both.* I shall eventually return to examine this question again in section 5.5.8 below; for the present I shall continue to take the former view, which regards `FullSR` simply as realising a limited form of genetic self-reproduction. The crucial feature of `FullSR`, for my purposes (i.e. in terms of P_a), is that it seems to be *robustly* self-reproducing, and, in this respect, it might represent a substantive advance over the various A-reproducers previously considered in Chapter 4.

As it happens, the reproduction mechanism of `FullSR` is also *genetic*—i.e. it is of the same general kind as formulated by von Neumann in the solution of P_v. In particular, we can roughly interpret the set of `E-OPs` in `FullSR` as a basic *Genetic Machine* (GM) g_0 as introduced in the previous chapter; the set of codon structures then collectively constitute a particular (dashed) A-descriptor, namely $d'(g_0)$, and self-reproduction follows. Indeed, although Holland does not explicitly mention von Neumann's work, it seems likely that Holland's particular definition of the `E-OPs` in α_0 was motivated precisely with this outcome in mind.

However, it should be emphasised that the similarities between α_0 and the (A-)systems introduced by von Neumann are very limited. While it is true that one can formulate an indefinitely large set of self-reproducing complexes, all "related" to `FullSR`, this has no particular significance in the context of α_0. Because the set of `E-OPs` defined in α_0 has been (deliberately) impoverished (to facilitate the analysis of this "proof of principle" system), the range of *A-behaviours* spanned by this set of A-reproducers seems to be extremely limited. In particular, since α_0 does not support universal computation, we cannot even expect this set of A-reproducers to meet the weak (behaviour spanning) criterion of allowing

the embedding of an arbitrarily programmed universal logical machine (ULM).

Thus, we may say that α_0 does indeed support a form of genetic self-reproduction, leading to an indefinitely large set of related A-reproducers (though even here, the details would diverge somewhat from von Neumann's concept), but this set of A-reproducers clearly does *not* span a significant range of complexity. α_0 would *not* therefore serve as a vehicle for the solution of P_v (by von Neumann's schema or otherwise).

While it is well to be clear about this divergence between α_0 and the von Neumann A-systems, it is not particularly surprising. The *problem* being addressed by Holland is (ostensibly) quite different from that tackled by von Neumann. Indeed, to the extent that von Neumann did, indeed, solve his identified problem, it is neither here nor there whether α_0 might provide another "alternative" solution to that same problem. No, the point of α_0 (for Holland at least) is not to consider the potential for evolution from `FullSR` to *more* complex A-reproducers, but rather to consider how even this initial, extremely basic, A-reproducer could itself arise from *less* complex precursors. In particular, Holland estimates that if the complex `FullSR` had to spontaneously emerge *solely* as a result of the unbiased generation due to the primitive operators, the expected emergence time would be of the order of 10^{43} time steps. Holland comments:

> This is such a large number that, for all practical purposes, we can reject the possibility of spontaneous emergence, if indeed the system [`FullSR`] must emerge in one fell swoop.
>
> Holland (1976, p. 399)

We now note that, on the face of it, `FullSR` itself is already of greater "complexity" (or, at least, bigger) than is strictly necessary. It would seem that we could shorten the arguments of the `E-OPs` of `FullSR`, while still retaining self-reproduction. Thus, shortening the arguments by one atom yields a complex of the form:

$$\begin{aligned}
&\texttt{-0:00-} && -N_0N_0N_1N_0N_0N_0N_0N_0- \\
&\texttt{-0:01-} && -N_0N_0N_1N_0N_0N_0N_0N_1- \\
&\texttt{-1:00-} && -N_0N_1N_1N_0N_0N_0N_0N_0- \\
&\texttt{-1:01-} && -N_0N_1N_1N_0N_0N_0N_0N_1-
\end{aligned}$$

Or, even more dramatically, if we shorten the arguments by a further atom (thus reducing them to the minimum length of just a single atom in each case),

it turns out that only one distinct argument (0) is required, and the complex can be reduced to just four distinct structures in total:

$$-\texttt{0:0}- \quad -N_0N_0N_1N_0N_0N_0-$$
$$-\texttt{1:0}- \quad -N_0N_1N_1N_0N_0N_0-$$

These simplified complexes share with `FullSR` the fact that they each define a set of possible transformations which, if they all occur, will result in the reproduction of the original complex.

But the issue here is not just which transformations are possible, but also which will actually occur. It should be clear that the simplifications of `FullSR`, suggested above, may be counterproductive: by shortening the arguments of the `E-OP`s one is making it more likely that `E-OP`s will bind with "random" codon strings (i.e. *not* belonging to the complex). If such "mis"-binding events are too common, then the complex will fail to achieve self-reproduction after all. In the event, Holland argues that `FullSR` represents a *minimal* complex in α_0 which could effectively self-reproduce.

However Holland goes on to identify complexes, significantly simpler than `FullSR`, which are not properly capable of self-reproduction (in the manner of `FullSR`) but which could nonetheless achieve a kind of "partial" (self?)-reproduction; he then argues that this phenomenon might be sufficient to strongly bias the subsequent generation of new structures (and complexes), and might ultimately provide a plausible route for the emergence of `FullSR` proper (for example).

Holland introduces firstly the following complex, consisting of three structures:

$$-\texttt{0:100}- \quad -N_1N_0N_0N_1N_0N_0N_0N_0-$$
$$-\texttt{1:100}-$$

The key point here is that the single codon structure in this complex, which can be copied and decoded by the two given `E-OP`s, does not code completely for either `E-OP`, but does code *partially* for *both*:

$$\gamma^{-1}(N_1N_0N_0N_1N_0N_0N_0N_0) = \texttt{:100}$$

The arguments of the two `E-OP`s are still long enough (arguably) to ensure that they will almost certainly bind to this codon string if it is available. So, if

this complex should arise, it should result in the generation of a high density of copies of the codon structure, *plus* a high density of E-OP fragments of the form:

$$-:100-$$

Now these fragments can be transformed into one or the other of the E-OPs of the original complex, simply by a B-atom being added in front of the colon atom—and it seems possible, at least, that this could happen spontaneously, with reasonable frequency, just by the background operation of the EX operator. This would effectively complete the reproduction of the original complex.[16]

Holland suggests that, indeed, this kind of process could occur and sustain itself in α_0. Indeed, he goes further and suggests that, if a large density of the relevant codon structure could be *initially* established, then this kind of process could actually be effective even if the arguments to the E-OPs were reduced from three atoms to just two. Again, this would allow the codon structure itself to be shorter also, so we identify the following complex as also "partially" self-reproducing:

$$-0{:}10- \quad -N_1N_0N_0N_1N_0N_0-$$
$$-1{:}10-$$

I shall refer to this complex as PartSR.

It is important to note carefully here the condition which allows the reduction in size of the E-OP arguments—namely that the complex (or, at least, the single codon structure within it) *already* exists in high density. More specifically, the claim is that if, by whatever means, instances of this codon structure, with strong internal bonding, should achieve high density, then a high density of the complex PartSR should spontaneously form and sustain itself indefinitely thereafter.

In more detail, the idea is that the initial large population of codon structures is expected to persist long enough that it is likely that an instance of the DC E-OP of PartSR will spontaneously form, even while the density of codon structures is still high. This will then result in the formation of a large population of the relevant E-OP fragments—i.e. fragments of the form:

$$-:10-$$

[16]It is perhaps debatable whether this should still be termed "self" reproduction—but the precise terminology is not important here. More generally, it does not matter for my purposes which, if any, of the complexes described are actually labelled as "living".

After this, the E-OP fragments could get spontaneously transformed into the required E-OPs belonging to PartSR, with sufficient frequency that a large population of PartSR does, indeed, form and sustain itself.

All this leaves open the question of how a high density of instances of the PartSR codon structure (with strong internal bonding) could be formed in the first place. To this end, Holland finally directs attention to the following complex with just two structures:

$$-0{:}100- \quad -N_1N_0N_0N_1N_0N_0-$$

Holland argues that if even a single instance of this complex should spontaneously form (with both structures having strong internal bonding) then this will result precisely in the formation of a high density of the PartSR codon structure, as required. Note the CP E-OP in this final complex is not the CP E-OP of the PartSR complex as such: it has an extra atom in the argument to ensure that it will preferentially bind, with "sufficient" probability, to the PartSR codon structure, even while the latter is present only at low density.

This final complex will be referred to as the Seed complex: it apparently has the property that, if a single instance should (spontaneously) form, then (with high probability) a population of PartSR complexes should arise, and subsequently sustain itself indefinitely (unless and until it is displaced by some other complex—perhaps even FullSR— which is more efficient in its reproduction).

Holland's "proof-of-principle" can then be stated as follows. A naïve view of the origin of "life" (in α_0) would assume that FullSR (or something of comparable complexity) must spontaneously form purely from the "unbiased" generation of "random" structures by the primitive operators. But, in fact, the complex Seed would spontaneously form at a much earlier stage (since it is so much simpler than FullSR). Holland specifically estimates the expected emergence time for Seed as only about 4×10^8 time steps, which would make such emergence quite feasible. Once this occurs, the subsequent generation of structures would be strongly biased, in a way which could dramatically accelerate the emergence of FullSR. Indeed, as indicated above, the process subsequent to the spontaneous formation of Seed would *already* take on at least some of the flavour of Darwinian evolution.

It may be noted that there is no claim that the complexes **Seed**, **PartSR**, or even **FullSR**, are *unique* in the rôles they play here. There may well be other complexes which, if they should spontaneously form, would strongly bias the subsequent generation of structures in a manner similar to that of **Seed** and **PartSR**, such that **FullSR**, or some other similarly complex A-reproducer, could then quite plausibly emerge. Holland's point is to give a "proof-of-principle": for this it is sufficient to exhibit *one* family of complexes (namely **Seed**, **PartSR** and **FullSR**) having the required properties. To whatever extent (if any) other alternative complexes could have an equivalent effect, Holland's argument could only be strengthened.

In this section I have been concerned solely with outlining the *conclusions* of Holland's analysis of α_0. This is necessarily qualitative—and entirely unconvincing in itself. Holland, of course, accompanies this discussion with a detailed quantitative analysis to support his conclusions. I shall not consider this analysis at this point. Rather, I shall turn to the more direct approach: simply testing whether the phenomena which have been qualitatively described here do, in fact, occur in a particular implementation of α_0.

5.4 AV0: A Realisation of α_0

This section describes a package called **AV0**, which is, in effect, a computer based realisation of α_0.

AV0 has been written entirely in the C language (ANSI X3J11). Original development was carried out on an IBM PC compatible platform, running MS-DOS, and Turbo-C V2.0. However, the empirical results reported below were recorded with an alternative version, still running on an IBM PC compatible platform, but compiled under **GNU cc**, and executed in 80386 protected mode, to allow access to a large, linear, 32-bit address space; the latter was required to allow universes of size greater than about 2×10^3 atoms to be realised. As far as possible, the package has been written to be "easily" portable (machine dependencies are encapsulated by conditional compilation). The source code comprises about 5000 lines, in roughly 60 files. This source code has been placed

in the public domain, and is available to interested researchers.[17]

The primary documentation for AV0 is the source code itself. This section is intended only to provide background information which might significantly ease the understanding of the source code.

5.4.1 The Programs

AV0 is organised into four executable programs.

av0run realises the α_0 dynamics proper. It offers facilities for loading and saving disk file images of α_0 universes and for executing the α_0 dynamic operators over any specified number of time steps, including dynamic display of a window onto the state string. There is also support for the extraction and logging of various statistical measures evaluated on the state string.

The other three programs are utilities for generating state strings with particular characteristics, as follows:

- **randmat**: This yields a completely "randomised" state string.
- **partsr**: This divides all material atoms more or less equally between instances of the three structures making up the PartSR complex, and free atoms.
- **fullsr**: This divides all material atoms more or less equally between instances of the eight structures making up the FullSR complex, and free atoms.

5.4.2 The Disk Images

A disk image of a particular α_0 minimally comprises three files, grouped by having the same name (up to 8 characters). The three files are distinguished by their extensions as follows:

- **.mat**: This contains an image of the state string.
- **.prm**: This contains all parameters not implicit in the state string—namely r, λ, m_1 and m_2. It also contains a set of flags which allow each of the

[17] Requests should be directed to the author, in the first instance.

operator types (`BM`, `EX`, `CP`, `DC`) to be selectively enabled or disabled. Finally, this file contains parameters specifying the interval (in time steps) at which log records should be emitted, and (separately) the interval at which the console display should be updated.

- `.stt`: This contains three "state" variables not contained in the state string: the current "time" (in α_0, not wall clock time), the current "seed" for the pseudo-random number generator, and a "count" of the number of pseudo-random number evaluations so far carried out (the latter is maintained as a check against possible cycling of the generator).

When `av0run` is executing it maintains a log file with the extension `.log`. Each log record contains a summary of certain statistics on the state string at a particular time step. When an image is loaded by `av0run` a log file will be created if one does not already exist; otherwise the new log records will be simply appended to the existing log file.

All disk files associated with the image of a particular α_0 are simple ASCII encoded text, so that they can be examined (and modified, if necessary) using normal text editing tools.

5.4.3 The State String

The primary data structure required is the state string. A closed doubly linked list is used. The size is dynamically determined whenever a new image is loaded, but remains static otherwise. This corresponds to a finite, but unbounded (circular) organisation, where the size is set at "initialisation". Essentially, the size is determined by an argument to whichever program is used to generate a `.mat` file; thereafter it is a constant of the α_0. The linked list is effectively superimposed on a simple, static, array of atoms.

A closed organisation was chosen primarily to avoid having to introduce special code to deal with behaviour at boundaries. Note that this means that the state string lacks any "absolute" position reference.

The linked list arrangement was chosen so that the locations *in memory* of each atom would not change (the relative locations in the state string do, of course, change). This makes the implementation of "movement" (which arises

both from the EX operator, and the E-OPs) reasonably efficient: a segment can be arbitrarily relocated just by rearranging pointers, rather than actually moving atomic symbols around in memory.

The underlying, static, array organisation makes it possible to efficiently maintain indices of specified kinds of atoms, which, in turn, can significantly speed up the execution of certain operators. However, many aspects of the α_0 dynamics still require segments to be scanned, generally in either direction. The double linking of the list makes this reasonably efficient.

5.4.4 Pseudo-random Number Generator

The AVO package involves a number of "stochastic" processes. A pseudo-random number generator is used to support these. A variety of pseudo-random number generators are available—there is generally one included with the standard C library. However, as reported by Park & Miller (1988), the quality of these generators is highly variable—where "quality" reflects some statistical measure(s) on the generated numbers. Furthermore, these statistical properties are not generally documented for the generators supplied as standard in a C library.

It was therefore decided to implement the "minimal standard generator", identified in (Park & Miller 1988), whose characteristics would, at least, be known, and would also meet certain minimal quality criteria.

5.4.5 Primitive Operators

Prima facie, the BM and EX operators both involve sampling a "large" number of independent Bernoulli random variables at each time step. This would be computationally expensive, and an alternative, statistically equivalent, algorithm was developed.

The key point is that, in each case, the probability of success for each Bernoulli random variable is quite small (typically 10^{-4} in the case of BM, and 10^{-2} in the case of EX). That is, "most" of the Bernoulli trials would usually come up with failure. The approach adopted was to first decide *how many* successes there should be on each time step, and then select *which* members of the population (of trials) will actually be the successes.

More formally, let $\mathbf{X}$ be an n-dimensional random variable, consisting of n identical, independent, Bernoulli random variables.[18] Interpreting a component value of 1 as indicating that the corresponding object should be operated on, and 0 as indicating that it should not, then the selection processes associated with the BM and EX operators are equivalent to making trials of a suitable $\mathbf{X}$.

Let $\mathbf{x}$ denote a trial of $\mathbf{X}$. Let M denote the number of 1's in $\mathbf{X}$. M, being a function of the random variable $\mathbf{X}$, is, formally, another random variable, which is jointly distributed with $\mathbf{X}$. Let m denote the number of 1's in $\mathbf{x}$—i.e. m denotes the result of the trial of M (note that, by the definitions of $\mathbf{X}$ and M, M is binomially distributed, with parameters n and p; this will prove significant later).

Let $q = 1 - p$. Given that the components of $\mathbf{X}$ are identical, independent, Bernoulli random variables with parameter p, the (marginal) probability function for any component is simply:

$$\begin{aligned} P(X_i = 1) &= p \\ P(X_i = 0) &= q \end{aligned}$$

Given that the components of $\mathbf{X}$ are independent, the probability function of $\mathbf{X}$ (i.e. the joint probability function of the components) is given simply by the product of the marginals. The event $\mathbf{x}$ denotes the situation that, of n independent Bernoulli random variables, m resulted in the value 1 (with probability p in each case) and $n-m$ resulted in the value 0 (with probability q in each case). The probability of this event is the product of the separate probabilities. Concisely, the probability function for $\mathbf{X}$ is, therefore:

$$p_{\mathbf{X}}(\mathbf{x}) = p^m q^{n-m}$$

This, then, is the objective: we wish to formulate an alternative procedure for evaluating $\mathbf{X}$, such that the probability function still matches this expression but which will be computationally more efficient (at least in the cases of interest in the α-Universe) than evaluating n independent Bernoulli random variables.

Now consider a new random variable, $\mathbf{X}'$. As with $\mathbf{X}$, $\mathbf{X}'$ is an n-dimensional random variable, where the sample space for each component is $\{0, 1\}$. $\mathbf{X}'$ is evaluated as follows. Let M' be a binomial random variable with parameters n

[18]The notation introduced here follows that of Larson (1974).

and p. Make a trial of M'; let m be the result. Randomly select m distinct numbers from the set $\{0..(n-1)\}$ (with all possible such selections being equally likely). For all i in this set assign the value 1 to the corresponding components, X'_i, of $\mathbf{X}'$; assign the value 0 to all other components of $\mathbf{X}'$.

Note carefully the distinction between M and M'. Though they both have the same probability function, M is defined (and therefore evaluated) *indirectly*—as a function of $\mathbf{X}$; whereas M' is directly defined (and evaluated) in its own right. The precise mechanism for directly evaluating M' will be discussed subsequently; for the moment, the important point is that a binomial random variable *can* be directly generated—we do not *have* to resort to the indirect method of generating n independent trials of a Bernoulli random variable.

I claim that, with this new procedure, $\mathbf{X}'$ will have precisely the same probability function as $\mathbf{X}$ (and may therefore be evaluated in place of it), but is computationally much more efficient (the quantitative improvement in efficiency will depend on p: the smaller p is, the greater the improvement). Informally, the idea is this: with the original procedure, we *always* had to evaluate n independent random variables—even though, "on average", only np of them resulted in a "success". With the new procedure, the number of random variable evaluations is, on average, $1+np$ (M' is always evaluated, and then, on average, a further np evaluations are necessary to randomly select the "lucky" components).

We now prove that this new procedure does indeed produce the same probability function as the original procedure.

M' is defined to be binomial, with parameters n and p; its probability function is therefore:

$$p_{M'}(m) = \binom{n}{m} p^m q^{n-m}$$

Now consider the event $\mathbf{X}' = \mathbf{x}$, where $\mathbf{x}$ contains exactly m 1's. This can occur only if, firstly, the result of evaluating M' is m (probability $p_{M'}(m)$ per the expression above), *and* the "correct" m elements of the set $\{0..(n-1)\}$ are selected. For the latter event, there are $\binom{n}{m}$ possible distinct outcomes, all equally likely—so the probability of the particular outcome specified is just the reciprocal of this. Since the two experiments (sampling M', and selecting the m

components) are defined to be independent, we can multiply the probabilities of the two events to get the probability of $\mathbf{x}$:

$$\begin{aligned} p_{\mathbf{X}'}(\mathbf{x}) &= p_{M'}(m) \cdot \frac{1}{\binom{n}{m}} \\ &= \frac{\binom{n}{m} p^m q^{n-m}}{\binom{n}{m}} \\ &= p^m q^{n-m} \\ &= p_{\mathbf{X}}(\mathbf{x}) \quad \text{QED} \end{aligned}$$

Assuming that the computational cost of evaluating a single (scalar) random variable is (approximately) independent of its probability function (and this assumption may have to be justified), then, clearly, it is computationally cheaper to evaluate only $1 + np$ random variables, instead of n. To quantify this, the new procedure will be, on average, computationally cheaper by a factor $n/(1 + np)$. In the case of `AVO`, we typically have $p < 10^{-2}$, and the computational advantage is substantial—say $> 10^2$.

In implementing this procedure in `AVO`, there are two distinct steps: evaluating M' (yielding m, the *number* of components to be selected) and then actually choosing the particular m components.

To sample a random variable of given probability function, a (pseudo-)random number is generated (with a uniform probability function over a given range) and this is then passed through an inverse, cumulative, version of the desired probability function. This distorts the uniform probability function of the original (pseudo-)random variable into just the shape of the desired probability function. Thus, the assumption that the cost of evaluating a random variable is independent of its probability function reduces to an assumption that the cost of evaluating the inverse, cumulative, probability function is negligible (at least compared to the cost of generating the pseudo-random number itself).

In the particular case of interest (evaluating M'—binomially distributed with parameters n and p) the computation of the inverse, cumulative function is quite

demanding, but this is overcome by using look up tables which are computed once-off per run of `av0run`, and imposes no on-going computational costs per time step. There is one lingering difficulty which is that the n values relevant to the various α_0 operators vary dynamically—they correspond essentially to the number of atoms having bonds of an appropriate sort (only atoms with weak bonds qualify as potential pivots for `EX` etc.). In practice this is overcome by fixing $n = R$ in *all* cases; then, m atoms are "provisionally" selected, but the operator is applied only to those for which it is allowed. This means that somewhat more atoms are "provisionally" considered for the application of the operator than is strictly necessary; but the net gain in computational efficiency is still judged to be worthwhile. The probability functions for selection of the "eligible" atoms are not altered by this procedure.[19]

The final issue here is, having generated a trial m of M, how to pick the appropriate m atoms. This is done in practice by repeatedly picking a random value in the range $0..(R-1)$ and using this as an index into the state string (viewed now as an array rather than a linked list). In itself, this runs the risk that a single atom could be selected more than once for a given operation within a single time step (it represents selection with replacement, rather than without replacement as required). This problem is overcome by tagging each atom which has been selected (for the given operation, on this time step); if such an atom is, by chance, reselected, that reselection is discarded, and another attempt made etc. Obviously, this process could become very inefficient if m could be comparable to R—but this does not arise in the cases of interest in `AV0`. The tagging process is made efficient by using a time stamp rather than a simple binary tag: this obviates the need to clear the tags on each time step.

[19]There is one minor qualification of this. The two distinct phases of `BM` (strong-to-weak and weak-to-strong) are executed sequentially. To keep the respective probability functions precisely as described by Holland it is technically necessary to ensure that a bond made weak in the first phase is not made strong again in the second phase of the same time step (albeit, with the typical parameter values, such an event would be extremely rare anyway). This is achieved by tagging each bond which is actually modified during the strong-to-weak phase; this tag can then be checked during the weak-to-strong phase. This tag is qualified by a (strong-to-weak) time stamp, so that it need not be explicitly cleared on each time step.

5.4.6 Emergent Operators

The simplest approach to the implementation of the emergent operators is to scan the entire state string, identifying (and processing) any emergent operators encountered. This is, indeed, the basic approach adopted in AV0. However, this scheme is modified in two ways, in order to improve the execution speed.

First note that, from the definition of the E-OPs, they all necessarily incorporate at least one colon atom. Now, typically, only one in 10 atoms in AV0 are colon atoms. Thus, instead of scanning the entire state string on each time step it is significantly more efficient to selectively target the scan onto the colon atoms. This is achieved by creating an index of all colon atoms; this need only be done once per run, at initialisation, as the "positions" of these atoms, in the *array* view of the state string, will not change thereafter. Then, on each time step, it is sufficient to inspect the immediate neighbourhood of each colon atom, in turn, to establish whether it is part of an E-OP. Again, a time stamp mechanism is used to ensure that a given E-OP (which may contain more than one colon atom) is not processed multiple times on any single time step.

The second, and ultimately more significant, optimisation is concerned with the location procedure for raw materials, which is an essential part of the execution of all E-OPs. Recall that there is no limit on how far this search may extend, possibly spanning the entire state string. The lower the density of the required raw materials, the more extended these searches will become. It turns out, for reasons to be explained later, that the densities of free atoms, particularly colon atoms, tend to quickly become quite small, and can be zero for a significant fraction of the time. Clearly, if there are no free atoms of a particular kind present in the state string at all, then the location procedure (for such a free atom) is guaranteed to fail; and should not even be initiated. This is arranged by establishing and maintaining counts of the free atoms of each element currently present in the state string. These must be correctly updated by the EX operator, and all E-OPs. For experiments in which the density of E-OPs is high, this optimisation has been found to yield approximately a five-fold increase in execution speed; in other cases the improvement is less dramatic, but is still worthwhile.

5.4.7 Tracking Complexes

In the experiments to be described it is essential to track the densities, in the state string, of specified complexes (specifically of `PartSR` and `FullSR`). In fact, it turns out to be sufficient to track an upper bound on these densities, which proves somewhat simpler to implement.

A first point to note is that, instead of literally attempting to assess the density of a complete complex, we track only the density of some one structure in the complex; clearly this does establish an upper bound on the density of the entire complex.

Secondly, note that we must decide *which* structure to track for any given complex. For both complexes of interest here, codon structures have been (arbitrarily) selected for tracking. In the case of `PartSR` the complex only includes a single codon structure ($-N_1N_0N_0N_1N_0N_0-$) so no further choice is necessary; in the case of `FullSR` the particular codon tracked was arbitrarily selected as $-N_0N_0N_1N_0N_0N_0N_0N_0N_0N_1-$.

However, some care must still be taken in identifying instances of these codon structures. Since codon structures are dynamically incorporated into the `E-OPs` and are punctuated by colon atoms (used as position markers), the tracking algorithm is designed to recognise the codon strings regardless of surrounding context, and regardless of the presence of a (single) embedded colon atom. Again, while this is not a completely reliable procedure, it will clearly yield a satisfactory upper bound for the density of the codon strings, and thus of the complexes, of interest.

5.5 Playing God

5.5.1 The Predictions

To recap, there are three substantive elements to Holland's predictions:

1. The `Seed` complex will spontaneously appear within a relatively short time (of the order of 10^9 time steps).

2. Once the `Seed` complex *does* appear, a population of `PartSR` complexes will be established, and will maintain themselves.

3. Conventional Darwinian evolution can then optimise the reproducing ability of the complexes quite quickly, up to and including the possible emergence of the `FullSR` complex (or something comparable).

Of these, the first potentially requires a substantial amount of (real) time to test; the second can be easily tested (by "playing God"—directly inserting an instance of the `Seed` complex); and the third can be tested only when (or if) testing of the second has been successful (i.e. after prediction 2 has been verified). Therefore, testing concentrated, in the first instance, on prediction 2—whether the `Seed` complex can establish a viable population of `PartSR` complexes.

5.5.2 Parameter Values

As already discussed in the detailed definition of α_0, Holland (1976) stipulated particular values for the α_0 parameters, which he then used in his numerical calculations. In summary, these values are as follows:

$$\begin{aligned} R &= 10^4 \\ \rho &= 0.5 \ (\Rightarrow \rho(-) = 0.5) \\ \rho(0) &= \rho(1) = \rho(:) = \rho(N_0) = \rho(N_1) = 0.1 \\ r &= 10^{-4} \\ \lambda &= 7/3 \\ m_1 &= 10^{-2} \end{aligned}$$

With minor exceptions, which will be noted below, these values were consistently adhered to in all the experiments to be described.

Holland did not specify any numerical value for the parameter m_2; in his Lemma 2 he suggests that his results will be insensitive to its exact value provided only that $m_2 < 1/b$ where b is the number of weak internal bonds in the structure(s) of interest. The longest structures to which Holland subsequently even loosely applies this analysis are the codons of the `FullSR` complex, which are each 10 atoms long, and thus have no more than 9 weak internal bonds. The condition $m_2 < 1/b$ is thus guaranteed to be satisfied, in all relevant cases, if we have $m_2 < 1/9$. It is desirable not to make m_2 much smaller than necessary, as this progressively slows the execution of the `EX` operator, and of `E-OPs` in general. Bearing this in mind, a value of $m_2 = 0.1$ in used in all the experiments described below.

I should note here that, in any case, I have been unable to follow the derivation of Holland's Lemma 2; and that I have carried out both theoretical and empirical analyses which suggest that it may be mistaken in detail. However, this seems to be a relatively minor issue, which would not critically affect Holland's predictions; therefore it will not be pursued further.

Four distinct experiments will be described. In each case, results are presented for two distinct runs of `av0run`; these runs are distinguished only in that the pseudo-random number generator was seeded with a different value, both in the randomisation of the initial configuration and the actual execution of `av0run`. The distinct seeds were chosen such that there was no overlap in the sections of the pseudo-random number cycle traversed within each pair of corresponding runs; that is, the runs used completely *distinct* sequences of pseudo-random numbers. The only purpose of this procedure is to demonstrate that, in each case, the essential pattern of the results does not rely on any artefact of the particular pattern of pseudo-random numbers encountered.

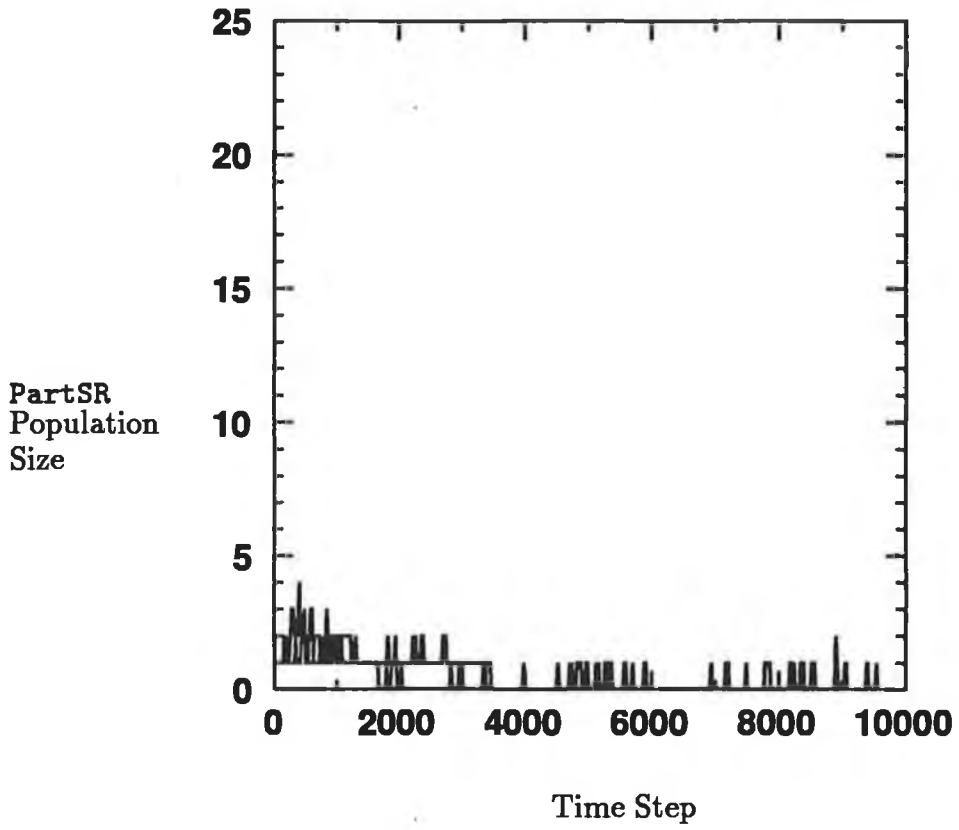

Figure 5.1: *The* **Seed** *Complex:* In this case, a randomised configuration is initially generated, and a single instance of the **Seed** complex is artificially inserted. The graph shows a superimposition of results from two runs, with disjoint initialisation of the pseudo-random number generator. Contrary to the original prediction, a large population of **PartSR** is *not* established.

5.5.3 Experiment 1: The Seed Complex

An α_0 of 10^4 cells was generated with a random initial configuration. A single instance of **Seed** was then inserted (this increases R slightly, and also slightly alters the element densities; but the effect is negligible). 10^4 steps were executed. This was repeated with the alternative seeding of the pseudo-random number generator.

The results of the two runs are shown (superimposed) in Figure 5.1. It shows the number of **PartSR** complexes present over time (as tracked by the number of **PartSR** codon structures). It is seen that, contrary to expectations, a significant population of **PartSR** was *not* generated. Indeed, the greatest **PartSR** codon string population achieved (itself an overestimate of the density of the complex proper) was only 4 instances.

5.5.4 Experiment 2: The *Modified* Seed Complex

On examining the detailed behaviour of the Seed complex it was found that, quite typically, the CP E-OP was *failing* to bind the PartSR codon, but was, rather, binding other "garbage" codons. But, reviewing Holland's analysis of Seed, we find a claim that the 3-atom argument of the CP E-OP *should* be sufficient "to assure that it will preferentially attach to the single copy of $\gamma(:\alpha(N_1N_0))$" (Holland 1976, p. 399).

In fact, it seems that Holland's analysis here is mistaken. Following the logic of Holland's own Lemma 1 (though not the precise result) the expected density of structures having the prefix 100 is $0.5 \times 0.1^3 = 5 \times 10^{-4}$. Thus, in a "region" of 10^4 atoms we would expect there to be approximately 5 distinct, garbage, codon structures which can be potentially bound by the Seed CP E-OP. Thus Seed could well fail to reliably reproduce the correct (PartSR) codon structure, and this indeed seems to be what is happening. Note that this effect is *not* offset by initially placing the two structures of the Seed complex immediately beside each other: with the relatively small value of m_2 in use (0.1), proximity has only a very limited effect on the binding probability.

To investigate this further, the previous experiment was repeated, but with the Seed complex *modified* by adding a further atom to the argument of the E-OP (the argument now becoming 1001). The expected number of garbage codon structures, matching this argument, in a universe of 10^4 atoms, now falls to 0.5, so the desired specificity should be achieved. It should be noted that this modification will also increase the expected emergence time of the modified Seed complex (though that is not, of course, at issue for these particular experiments); and that the expected lifetime of the modified Seed complex will be somewhat reduced thus reducing the maximum density of the PartSR complex which could be generated. But, for the moment, the important requirement is to establish a significant density even of the PartSR codon structure.

Figure 5.2 is a plot of the outcome of this experiment. While there *is* now a significant generation of PartSR complexes (or, at least, of its codon structure), it is clear that this effect is still very limited. The greatest PartSR codon string population achieved was 18 instances, whereas this particular universe has a

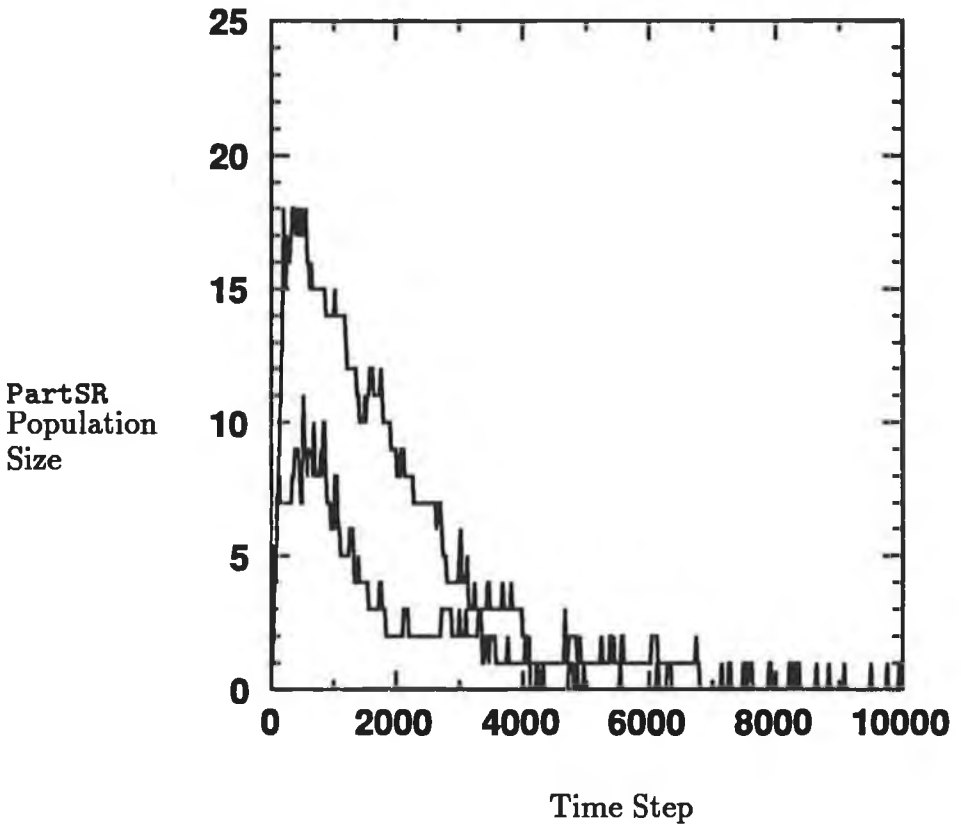

Figure 5.2: *The Modified* **Seed** *Complex:* In this case, a randomised configuration is initially generated, and a single instance of the *modified* **Seed** complex (see text) is artificially inserted. Again, the graph shows a superimposition of results from two runs, with disjoint initialisation of the pseudo-random number generator. While a population of the **PartSR** complex (or, at least, its codon string) *is* now initially built up, it subsequently dies out again relatively quickly.

theoretical capacity for 250 instances. It is clear that the modified **Seed** complex does not come close to saturating the universe in this sense. Furthermore, after the initial transient, the population rapidly dies out.

5.5.5 Experiment 3: The PartSR Complex

At this point it was clear that the **Seed** complex was not capable of carrying out the function anticipated by Holland—i.e. to establish a viable population of **PartSR** complexes. However, it was not clear whether this was merely a problem of the relatively limited size of **PartSR** population which the initial instance of **Seed** was managing to generate, or whether the **PartSR** complex would not be viable even in an established population.

To test this, an α_0 was generated (via the **partsr** program) with a highly artificial initial configuration—essentially (80%) saturated with instances of the

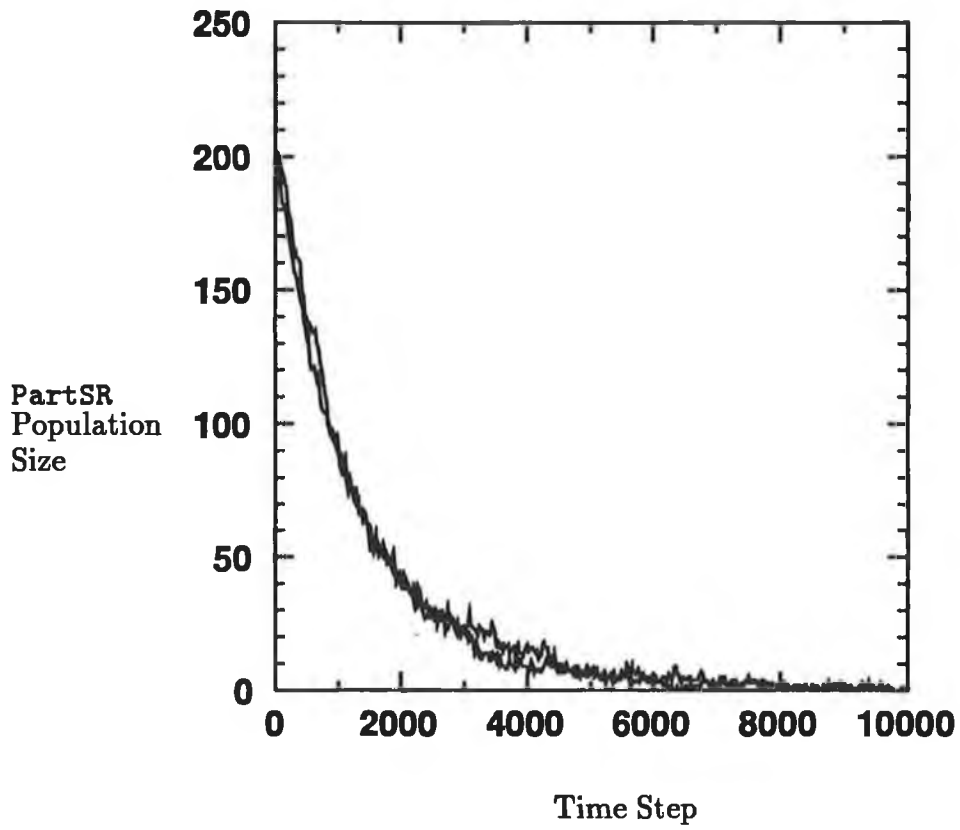

Figure 5.3: *The* `PartSR` *Complex:* In this case, an initial configuration is generated by the program `partsr`; this consists solely of instances of the structures making up the `PartSR` complex and of free atoms. Again, the graph shows a superimposition of results from two runs, with disjoint initialisation of the pseudo-random number generator. Contrary to the original prediction, the population of `PartSR` is *not* sustained, but dies out rapidly.

`PartSR` complex. This was again executed for 10^4 steps. Figure 5.3 shows the outcome. It is seen that, even with this "most favourable" configuration, the population still rapidly goes extinct.

5.5.6 Experiment 4: The `FullSR` Complex

It will be recalled that the `PartSR` complex has the property of not being "fully" self-reproducing: it relies on the primitive operators to complete its cycle of reproduction. A final experiment was carried out to test whether this was a critical factor in the failure of `PartSR` to be viable. In this case, an α_0 was generated (via the `fullsr` program) which was saturated with instances of the `FullSR` complex (the maximum capacity is 35 instances).[20] This was again executed for 10^4

[20]The size of this universe was made marginally larger than 10^4. A basic complex consisting of a single instance of `FullSR` plus sufficient free atoms to correctly set the relative element

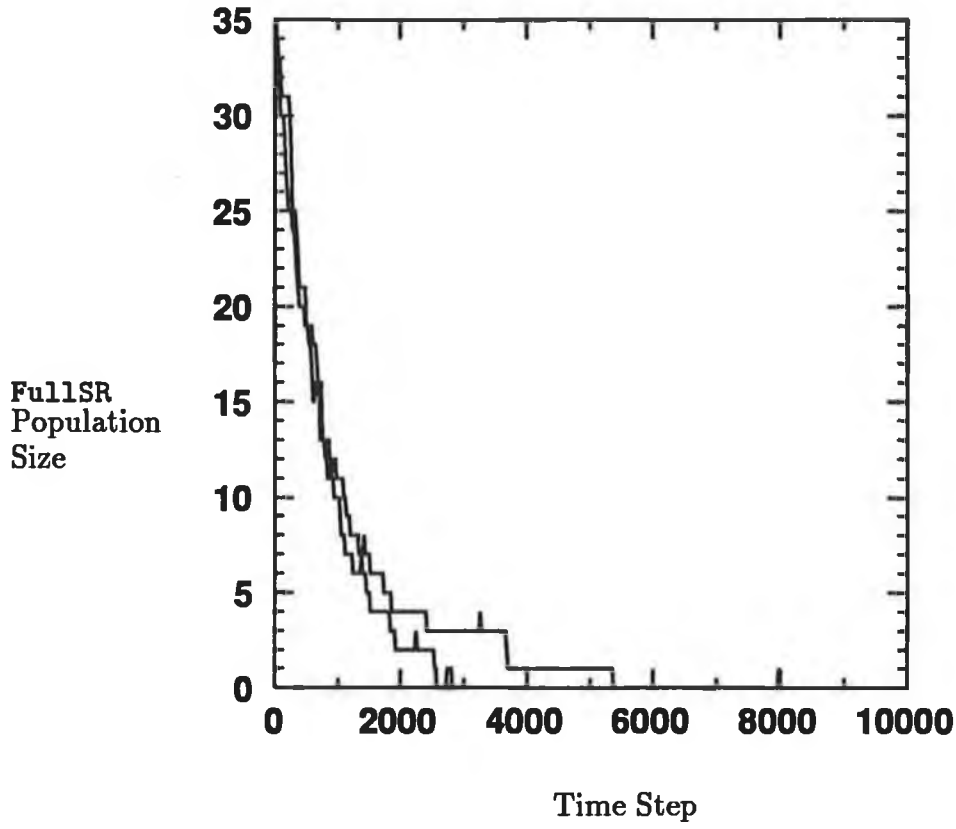

Figure 5.4: *The* `FullSR` *Complex:* In this case, an initial configuration is generated by the program `fullsr`; this consists solely of instances of the structures making up the `FullSR` complex and of free atoms. Again, the graph shows a superimposition of results from two runs, with disjoint initialisation of the pseudo-random number generator. As with the `PartSR` complex, and contrary to the original prediction, the population of `FullSR` is *not* sustained, but dies out rapidly.

steps. Figure 5.4 shows the outcome. It is seen that, even with this "fully" self-reproducing complex, the population still rapidly goes extinct; indeed, the extinction is, if anything, somewhat more rapid for this complex.

5.5.7 What's going wrong?

From the experiments described above, it was clear that α_0 was simply not capable of supporting the "life-like" behaviour postulated by Holland. There was thus no point in pursuing the question of spontaneous emergence. But these experiments do not, in themselves, give any indication of how deep rooted (or otherwise) the deficiencies of α_0 may be.

densities turns out to require 290 atoms. The size of the universe must then be made an integral multiple of this (10, 150) to correctly retain these densities.

A series of informal studies were then carried out, which involved simply monitoring a dynamic display of a window onto the α_0 state string, over many different variations in the configuration and parameters of the universe. Based on this exercise it was possible to identify at least some specific, proximate, causes of failure (there can, of course, be no guarantee that *all* relevant factors were identified by this process).

Note that Holland's analysis of the `PartSR` dynamics relies on its being composed of structures which are, internally, "strongly bonded"—which is to say *long lived.* He then estimates the average "productivity" over this lifetime, to come up with a net positive rate of change for the density of the complex (once a threshold is reached). However, in practice, there are (at least) three factors which severely disturb the behaviour of the complex, and which are not allowed for in Holland's analysis:

- Raw materials (free atoms, in particular) quickly become scarce (due to usage by random, garbage, emergent operators). The effect is to drastically reduce the rate at which all emergent operators function in practice, thus reducing the *fecundity* of any putatively self-reproducing complexes.

- Even when a structure is strongly bonded internally, there is nothing to stop random garbage moving into a position immediately adjacent to it. At the very least this interrupts or suspends the progress of an emergent operator. Thus, it turns out that complexes can only be *active* for a limited portion of their total lifetimes (regardless of the availability of free atoms); again this severely limits fecundity.

- But, at worst, this random arrival of a garbage structure beside an emergent operator can have much more severe effects. If it arrives on the right hand side it can corrupt the output of the operator (introducing a high "mutation" rate, and further reducing fecundity). If it arrives on the left hand side it can result in the formation of a different, garbage, emergent operator which forcibly, and prematurely, breaks up the original operator. This has actually been observed to occur on a number of occasions. Thus, as well as reduced fecundity, complexes also have higher mortality than expected.

So: compared to Holland's analysis, the lifetimes of the structures are shorter than expected, they are only active for a fraction of this time, and their products are quite frequently corrupted. The net effect is that mortality exceeds fecundity (by a significant margin), the putatively reproducing complexes cannot make up for their own natural decay rate, and thus become extinct quickly. These effects are directly related to the time required to complete a reproduction cycle, and thus to the size of the complexes. This explains the even worse performance of the `FullSR` complex compared to `PartSR`.

Note that I have not presented any experimental investigation of the sensitivity of my quantitative results (experiments 1–4 above) to the parameters of α_0, even though this would be a straightforward (if tedious) exercise. The reason for this omission should now be clear: Holland's analysis has been demonstrated to be greatly oversimplified, and defective as a result. The failure of the predictions is not dependent on the particular parameter values used, but is, rather, representative of the fact that several significant factors have been entirely neglected in the analysis. This effectively destroys the assumed theoretical basis for the empirical investigation; in this situation a random search through the α_0 parameter space for a system which would show some "interesting" behaviour seems to me quite futile, and I do not pursue it.

Thus, even though the original objective was to design an artificial system which *would* be simple enough to allow closed form analysis, it turns out that α_0 is just not that simple. At this point it seems doubtful to me that a system which was genuinely simple enough to allow the kind of analysis envisaged by Holland, would actually support any of the phenomena of relevance—though this remains an open question. I should note that Holland's own diagnosis of the situation is not quite as pessimistic as mine; commenting on the preliminary results of the current work (McMullin 1992d) he stated:

> When I wrote my original version I didn't make any allowance for diversionary "precipitates" (and the like) that would sequester needed intermediates ... several of my colleagues in various places have said in one way or another that this is a central problem in the origin of life. I think some revisions in the model would "fix this up" but it takes some thought (and I would make no guarantees).
>
> John Holland (personal communication)

In any case, leaving aside the question of closed form analysis, a naïve attempt at solution of the particular problems observed in α_0 might be to simply reduce the "temperature" of the universe (reduce the rate at which bonds decay, and structures get randomly moved around). It seems likely that the `FullSR` complex could be made "viable" in this way (in the limit, if the primitive operators are disabled entirely, `FullSR` should be able to expand to the capacity set by whatever free atoms are initially available; thereafter, of course, all dynamic activity would cease); however, this is not at all the case for `PartSR`, which relies on the primitive operators to complete its reproduction. It is quite possible (though no proof is currently available) that `PartSR` would not be viable at *any* "temperature". However, in any case, from the point of view of the problem originally posed by Holland (i.e. that of spontaneous emergence) any reduction in "temperature" would be accompanied by an increase in the expected emergence time for any particular structure or complex, and thus may be completely counterproductive.

But there is a more general point here: my purpose in studying α_0 was not concern for the problem of spontaneous emergence *per se*, but as an avenue to the solution of P_a—the demonstration of A-reproducers which are *robust* in the face of environmental perturbations (including interactions with each other). Thus, while one might well be able to improve the viability of certain complexes in α_0 by *ad hoc* measures which "protect" them from interference, this would be to undermine completely the purpose for which I turned to the α-Universes in the first place. The specific perturbations identified above, which arise in α_0, are precisely the kinds of things we *want* to allow. Again, as noted at the end of the previous chapter, we already *know* that we can achieve "viable" A-reproducers if we rule out, or rigidly constrain, their interactions with each other and their common environment, so modifications of α_0 which move in that direction are fundamentally of limited interest.

In summary, α_0 does *not* yet provide any substantive advance toward a solution of P_a. The A-reproducers in α_0, such as they are, are just as fragile as in any of the A-systems considered in the previous chapter; contrary to Holland's analysis, α_0 does not provide any prospect for the spontaneous emergence of *robust* A-reproducers, and does not, therefore, provide a basis for the realisation of Artificial Darwinism.

5.5.8 Can We Fix it?

I should note here that the original paper (Holland 1976) seems to have been largely ignored since its publication. I have been able to identify only two substantive discussions of it: by Martinez (1979) and Kampis (1991, Section 5.1.2). In both these cases the correctness of Holland's analysis was *assumed*, and further discussion was then predicated on that. Given the results which have been presented here, this assumption was not justified; I shall not, therefore, comment further on these works.

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

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

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

This is, of course, rather speculative; but, as it happens, it is closely related to a general model for the origin of *terrestrial* life which has been championed by Cairns-Smith (1982). This is based on *inorganic* information carriers, which could conceivably be replicated without the relatively complex apparatus required

for RNA or DNA replication. It seems to me, in the light of the experimental results presented here, that it would now be a promising research program to adopt Holland's original *strategy* (which is to design relatively simplified model chemistries, loosely based on cellular automata, in which to examine the origin of "life"), but to replace his detailed models (the α-Universes) with models based on different theoretical considerations—such as those of Cairns-Smith.

The second avenue I can envisage for challenging the limitations of α_0 turns on a point which is both subtle and fundamental. I had already raised, or at least anticipated, this issue in general terms in the previous chapter (section 4.3.4), and I referred to it again, albeit obliquely, earlier in the present chapter (section 5.3). In the specific context of α_0 it may be expressed in the form of a question: should complexes, such as `FullSR`, be properly regarded as realising self-reproduction (as I have done up to this) or, instead, as realising a primitive form of autopoiesis?

Briefly, the situation is this. As long as we consider an instance of an A-machine in α_0 as corresponding to a particular, fixed, set of structures, then it makes sense to regard the mutually recursive relations of production between these structures as realising a form of self-reproduction—such a set of structures is (in principle at least) capable of bringing new and separate instances of such sets into existence. But this is not the only possible way of looking at things. We could, instead, regard an A-machine in α_0 as corresponding to the set of recursive relations of production rather than a particular set of structures which happen to realise these relations. In effect, an A-machine is then identified with what I have previously regarded as a *population* of structures (or complexes) in α_0. These relations of production are then recognised as being autopoietic: such a population is (or, at least, should be) capable of sustaining itself, by virtue of this autopoietic organisation, despite turnover of some or all of its constituent structures.

From this perspective, the phenomena studied in α_0 can now be recognised as fundamentally related to phenomena occurring in, say, the `VENUS` (Rasmussen *et al.* 1990) or `Tierra` (Ray 1992) systems, discussed in the previous chapter (see sections 4.3.1 and 4.3.3). In the present chapter I have very loosely talked in terms of the putative A-reproducers in α_0 as being potentially "robust" or "viable"; but the fact is that, as long as by "A-reproducer" I meant a single fixed

set of structures, there was never any possibility of their being "autonomous" in the strong sense of being *autopoietic*. As it happens, the putative α_0 A-reproducers turned out not be "viable" anyway (just like the A-machine MICE in VENUS); but, even if they had been "viable", it seems that it could only have been, at best, the cosseted "viability" of the A-reproducers in Tierra with their inviolable memory allocations. By definition, no *static* set of structures in α_0 can realise the *dynamic* homeostasis of its own identity, which would be characteristic of properly autopoietic viability or autonomy.

By contrast, if we turn our attention to "populations" of structures in α_0—the equivalent of considering "organisms" in VENUS or "sociality" in Tierra (see Chapter 4, section 4.3.4)—we *can* encounter the possibility of properly autopoietic organisation. Granted, in α_0 as it stands, the autopoiesis is not effective—such populations actually die out—but (with the example of Tierra before us) we may anticipate that some modified α-Universe could overcome this. The point is that the kinds of entities which we might properly regard as autonomous are *not* the kinds of entities which could be regarded as self-reproducing; and, moreover, the "higher level", properly autonomous entities, are not, in general, self-reproducing in any sense, and are *certainly* not genetically self-reproducing in the von Neumann sense of permitting an open-ended growth in complexity.

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

Well, the first point is that to have *any* kind of self-reproduction, we would probably need some mechanism for the formation and maintenance of *boundaries* by the autopoietic entities. Some kind of boundary formation is actually part of the definition of fully fledged autopoiesis. Furthermore, a boundary seems to be logically necessary if we wish to talk about self-reproduction: unless the entities establish well defined boundaries then it is entirely unclear what could possibly qualify as self-reproduction. In VENUS, Tierra, or α_0, as they stand, there are no such mechanisms for boundary formation (capable of bounding the *relevant* entities). Boundary formation has, of course, been exhibited in the A-systems pioneered by Varela *et al.* (1974). These systems, by contrast to VENUS, Tierra and α_0, are *two* dimensional rather than linear. On the other hand, the

introduction of a kind of boundary mechanism has been previously outlined by Martinez (1979), in a modification of α_0 which would still be one-dimensional. Thus, while two-dimensionality is probably not essential here, it certainly provides conceptual simplification, and makes visualisation much easier.

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

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

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

Conversely, if we used α_0 as our starting point, and succeeded in modifying it to support reproduction of bounded, autopoietic, "populations", then we would have entities which *do* exhibit a "von Neumann style decoding"; but, of course,

they would be impoverished in a different manner, namely that the functionality available in α_0 is extremely impoverished anyway and there certainly could not exist a space of such autopoietic A-reproducers which would span a wide range of A-complexity (see the previous discussion of this point in section 5.3 above).

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

5.6 Conclusion

> The popular view that scientists proceed inexorably from well-established fact to well-established fact, never being influenced by any unproved conjecture, is quite mistaken. Provided it is made clear which are proved facts and which are conjectures no harm can result. Conjectures are of great importance since they suggest useful lines of research.
>
> Turing (1950, p. 442)

I should like to emphasise the debt which the work reported here owes to John Holland's original formulation and analysis of the problem of spontaneous emergence of self-reproducing behaviour. While it has been possible to point to defects in that analysis, this was with the benefit of hindsight, and prompted by experimental evidence not available to Holland. It does not detract in any way from Holland's creative achievement in formulating the *possibility* of such an investigation in the first place.

In conclusion, this chapter is a report on failure—but, I suggest, in the very best and most productive sense of that word. As enunciated by McDermott in the quotation with which I opened the chapter—and, indeed, as encapsulated in the Popperian theory of the evolutionary growth of knowledge—failure, or experimental refutation of predictions, is the very stuff of the so-called "scientific method". Although the model universe α_0 fails to demonstrate the phenomena originally hoped for, its particular mechanisms of failure are interesting in their own right. These show that P_a continues to be a deep and intractable problem *even* in a universe which is extremely simplified, and where the dynamics

have been deliberately tailored to make von Neumann style genetic reproduction "easy" to realise.

However: all human works are finite, the end of the rainbow steadily recedes, and we have finally reached that otherwise arbitrary point where a halt must be called to this cycle of conjecture and refutation. It only remains, in our concluding chapter, to briefly look back and consider a final, distinctive, view, which can be made available now that we have arrived at the end point of this particular journey.

Chapter 6

Rainbow's End?

> The way in which knowledge progresses, and especially our scientific knowledge, is by unjustified (and unjustifiable) anticipations, by guesses, by tentative solutions to our problems, by *conjectures*. These conjectures are controlled by criticism; that is, by attempted *refutations*, which include severely critical tests. They may survive these tests; but they can never be positively justified: they can be established neither as certainly true nor even as 'probable' (in the sense of the probability calculus). Criticism of our conjectures is of decisive importance: by bringing out our mistakes it makes us understand the difficulties of the problem which we are trying to solve. This is how we become better acquainted with our problem and able to propose more mature solutions: the very refutation of a theory—that is, of any serious tentative solution to our problem—is always a step forward that takes us nearer to the truth. And this is how we can learn from our mistakes.
>
> As we learn from our mistakes our knowledge grows, even though we may never know—that is, know for certain. Since our knowledge can grow, there can be no reason here for dispair of reason. And since we can never know for certain, there can be no authority here for any claim to authority, for conceit over our knowledge, or for smugness.
>
> Popper (1989, Preface to the First Edition, p. vii)

This quotation from Popper captures, perhaps, the single most important idea in all of Popperian philosophy. It certainly identifies the central, unifying, theme of this Thesis: in brief, I have tried to take this Popperian philosophy and methodology seriously, and to apply it in the context of Artificial Intelligence.

However, there have been some diversions and digressions along the way, so it may be as well to finally distil out the central ideas again. These may be loosely represented as a series of interrelated conjectures. I shall identify each in turn, and comment briefly on how I have dealt with them.

- Conjecture: *Mentality is computational.*

  This conjecture underlies and motivates much of AI; but it is deeply counterintuitive and even repugnant. I considered two separate substantive attempts to refute this conjecture—by Searle and by Popper—but concluded that they were flawed; this leaves the status of the conjecture open, and I tentatively adopted it.

- Conjecture: *Knowledge is computational.*

  This conjecture characterises AI in the "weak" sense, where we ask only that a computer system display "intelligent behaviour", without committing ourselves as to its "genuine mentality". In considering this conjecture, my primary concern was to clarify the interpretation of "knowledge"; I concluded that provided we mean something like "effective anticipation" then knowledge is, or at least can be, computational.

- Conjecture: *Computational Knowledge can grow.*

  In my view this conjecture epitomises the most difficult and fundamental challenge within AI. Having analysed it, my conclusion was that computational knowledge can indeed grow—but *only* by a process of unjustified variation and selective retention; so the challenge becomes to design computational systems which can realise such processes.

- Conjecture: *Artificial Darwinism is possible.*

  The point at issue here is whether artificial, computational, "knowledge" or "complexity" can grow by a process of Darwinian evolution. Von Neumann pointed out that there is a *prima facie* refutation of this: it seems paradoxical that any automaton could construct another of greater complexity than itself. Von Neumann went on to show how this argument is, in fact mistaken, and such growth of complexity can be supported by a form of genetically based self-reproduction. This, however, leaves the question of *autonomy*—which is also required for Darwinian evolution—open.

- Conjecture: *Artificial Genesis is possible.*

  One plausible route to achieving artificial Darwinism is to realise some form of artificial genesis of Darwinian actors. I examined one very specific elaboration of this conjecture, in the form of Holland's α-Universes; and concluded that the conjecture was refuted in that particular case, but the refutation was productive in suggesting some alternative reformulations.

Rather than review these various points in greater detail again, I shall try to finally conclude in a different way. The genesis of an idea is, of course, a different thing from its validity. In laying out this Thesis I have naturally tried to organise the material, with the considerable benefit of hindsight, into its most "logical" order; but this is certainly not the order in which it originated. In closing then, I should like to offer, briefly, that alternative perspective on the Thesis, which tries to show where the ideas actually came from and how they grew. I shall present it almost as an autobiographical record—but of course, the significance lies not in the World 2 of my personal subjective experiences as such, but in the additional insight which this narrative may yield into the World 3 problems which I have been concerned with.

I must start with the years from 1983 to 1987, which I spent working with *Hyster Automated Handling Limited* (HAHL), as an engineer and a manager, on the design of "Automatic Guided Vehicles" (AGVs)—in effect, a form of mobile robot—and systems thereof. I was privileged to work with an extraordinarily talented and enthusiastic team in HAHL over those four years; we started with the proverbial blank drawing board, and despite extreme youth and inexperience, we designed, built, and installed several successful AGV systems in Europe and North America.

Traditionally, AGVs had been designed to be "dumb": for the most part, their functionality was controlled by some kind of off-board controller, typically a large central computer. In HAHL we set out to design "intelligent" AGVs; later, we even went so far as to call them "autonomous". They were intended to operate, as far as possible, without relying on direction from any off-board, or central, controller. Our success in this was, of course, only partial, but it established an approach or objective which has since become standard in the industry.

I arrived in Dublin City University (then the National Institute for Higher Education, Dublin) in June 1987. In making this move I was specifically motivated by a desire to investigate some of the deeper, fundamental, problems of building genuinely "autonomous" systems. I was conscious of the fact that, despite our successes in HAHL, these vehicles were still, in truth, very stupid, if they were compared with even the simplest of biological organisms. While, from a practical, technological, point of view, the brute force method of trying to "engineer" smarter machines still seemed like the correct way forward, I was convinced that, in the longer term, more fundamental advances would be required.

Given this background, it was a small step to the idea of mechanising some kind of Darwinian process—after all, that was how biological organisms were "designed". I was not yet aware of the complexities underlying this idea!

My earliest investigations were concerned with the work by Holland (1986) on *Classifier Systems*, and Reeke & Edelman (1988) on *Neural Darwinism*; I felt that there were significant underlying parallels between these apparently separate developments (McMullin 1988). I gradually expanded outward to identify other workers who had formulated what appeared to be related approaches (McMullin 1989). I recognised a common core here, concerning the growth of computational knowledge through some kind of essentially recursive or self-referencing process. I dubbed this rather vague idea the *reflexive hypothesis*; I was conscious that there was a danger of paradox or infinite regress here, which I characterised by the question *Who Teaches the Teacher?* (McMullin 1990). I had stumbled—though I surely did not yet recognise it—on the problem of *induction.*

I am not sure when I first read Popper's *Objective Knowledge: An Evolutionary Approach* (Popper 1979); but I know that I returned to this marvellous collection of essays many times, and it provided immeasurable clarification for the whole enterprise.[1] More specifically, Popper provided me with a coherent account of how the regress implicit in the evolutionary growth of knowledge can be made benign rather than vicious, and this allowed me to examine the problem situation in Artificial Intelligence, and machine learning, with a quite new

[1]It is, of course, for this reason that I chose to play on Popper's title in naming this Thesis.

perspective. This ultimately produced the detailed discussion presented here in Chapter 3.

However, I was also conscious that Popper rejected physicalism in general, and computationalism in particular (Popper & Eccles 1977). If I was to continue with the methodology of computationalism, I needed to at least understand these views of Popper. Coincidentally, John Searle's rather different criticisms of computationalism were also receiving something of a revival at about this time (Searle 1990). I found that I was sympathetic with the intuitions being expressed by both Popper and Searle—computationalism is certainly not an *attractive* idea—but I could not accept that their arguments were remotely decisive. This critique of Searle and Popper became, in effect, Chapter 2 of the present work.

Somewhat in parallel with these developments, I was still working on the problem of realising a "satisfactory" form of Artificial Darwinism. I was not sure what I meant by "satisfactory", but I was sure that the things I had found in the literature (such as the "Genetic Algorithm" in particular) were not it. At this point it seemed to me that, if I wanted to achieve a spontaneous *growth* of knowledge or complexity, I might just as well ask for the spontaneous *emergence* or *genesis* of complexity. Informally, I wanted to reduce, or eliminate, the possibility that I, as the developer or programmer, would be directly or indirectly "injecting" complexity into the system; and it seemed that this constraint would *surely* be satisfied if the system were started in a totally "random" or "chaotic" state.

I then came upon Holland's description of the α-Universes, and, in particular, the system which I have called α_0, and which Holland analysed in detail (Holland 1976). While the functionality or "potential" for organisation that would be possible in α_0 was clearly extremely limited, it did seem like this could provide a good starting point for the kinds of system I wanted to investigate. Moreover, by this stage I wanted to tackle something more concrete. I found Holland's theoretical analysis extremely difficult to follow, and I therefore resolved to carry out an empirical investigation. That is, I would build α_0, and play with it. On the one hand, this would help me understand and probe Holland's theoretical results, and I would also then be in a position to decide how to enhance α_0 in an effort to achieve more substantive spontaneous organisation.

In the event of course, I discovered that α_0 was significantly more complicated in its behaviour than Holland had anticipated, and the predictions of his analysis did not hold up in practice. This rather stymied the idea of "simply" enhancing α_0. First indications of the negative results regarding α_0 were informally communicated at the AICS '89 conference, held in DCU in September 1989; following much more extensive testing, a concise published account eventually appeared as (McMullin 1992d). The fully detailed description of this work, including a *complete* formal specification of α_0 (which was neither required nor provided by Holland), now comprises, of course, the substantial part of Chapter 5 of this Thesis.

I was extremely dissatisfied with the outcome of the experiments on α_0, but was very unclear on how to possibly move beyond them. While Holland did not say so explicitly, it was clear that the kind of "genetic" self-reproduction which he had envisaged would emerge in α_0 had been inspired by John von Neumann's work on "self-reproducing automata" (Burks 1966d). I therefore resolved to study this original work by von Neumann carefully. This turned into a very prolonged exercise. Adopting a Popperian approach I tried to ask what problem(s) von Neumann had been trying to solve; and I found that the answers which seemed to be offered by Burks, and by various subsequent commentators, did not stand up to critical examination. In particular, it seemed to me that the idea of "universal construction" which von Neumann had formulated had been subsequently interpreted in a variety of different, and mutually contradictory, ways. Evidently something was amiss, but I was not at all sure just what.

What *was* clear to me was that von Neumann was concerned with the growth of "complexity" by essentially Darwinian means—and that he was only interested in self-reproduction as a means to this end. I conjectured that a von Neumann style "genetic" self-reproduction *might* be some kind of *necessary condition* for such Darwinian evolution, and that *this* was von Neumann's "result". I was strongly influenced in this view by the various writings of the evolutionary biologist Richard Dawkins, especially *The Selfish Gene* (Dawkins 1989b) and *Universal Darwinism* (Dawkins 1983). Dawkins seemed to have arrived at a similar conclusion in regard to the necessity of "genetic" self-reproduction, though by an entirely independent route.

There followed an interlude, during which I attempted an intensive study of at least a selected fragment of the literature of evolutionary biology, in an attempt to make sure that I properly understood Darwinian theory in its original setting. As a result I attempted to reformulate the theory in an entirely abstract form (McMullin 1992a), and then reviewed biological (or shall we say "organismic") Darwinism from this perspective (McMullin 1992b). Finally, and most importantly, I used this as a basis for an extensive and detailed critique of Dawkins' "genic selectionism", showing first of all that the presentations of it have sometimes been less than consistent, and secondly trying to separate out those elements which can be successfully defended (McMullin 1992c). I had originally intended that all this biological material would be integrated into the Thesis; but, in the event, it expanded to far too great a length, and was not essential to the understanding of the other material in any case; it was therefore separated out into the several technical reports just cited.

While this biological review no longer appears overtly in the text of the Thesis, it had a very important and necessary effect, nonetheless. It was only after completing this exercise that I was able to properly formulate the detailed analysis and re-interpretation of von Neumann's work which now appears as Chapter 4. Specifically, as long as I tentatively accepted Dawkins' doctrine of genic selectionism, I was not able to clearly envisage what problem von Neumann might have been attempting to solve. Contrariwise, once I had satisfied myself that genic selectionism, in Dawkins' terms, could be validly rejected (or, at least, radically diluted), I was free to recognise von Neumann's true achievement: this was to show, not that genetic self-reproduction is a *necessary* aspect of Darwinian evolution, but that it is one *possible* means of allowing such evolution. That is, von Neumann's problem was to show how a spontaneous growth of complexity could be possible *at all* in a mechanistic world.

With this resolution of my doubts about von Neumann's work, it was time to return again to the α-Universes, and the question of spontaneous emergence of von Neumann style self-reproducing automata. Having once established what problem von Neumann *had* solved, it became clear, by omission, what was outstanding—and, indeed, what should be sought from any revised or enhanced version of α_0. I initially expressed this in terms of words like "robustness" and

"viability", and the connection between these things and the possibility of natural selection. It was with the benefit of this idea that I criticised the more recently published attempts at artificial Darwinism, such as `VENUS` (Rasmussen *et al.* 1990) and `Tierra` (Ray 1992). But it was only when I discovered the notion of autonomy in the technical sense of *autopoiesis* (Maturana & Varela 1980), that the final element fell into place. It was this that gave me the concepts and vocabulary which allowed me to properly complete the discussion of α_0, `VENUS` and `Tierra`, and to draw out the fundamental similarities between these superficially diverse systems, and to identify the prospects for a future synthesis.

With this very late addition, the Thesis was finally completed; or at least as complete as any such work ever can be.

And as to the end of the rainbow? I do not know now whether, as a child, I really believed that I could get there; or if, having arrived, I would find the unfortunate Leprechaun's crock of gold, and quickly, quietly, steal it away. But though I chased many rainbows then, and since, I did gradually realise that the fun was in the chase, and in the beauty of the rainbow itself.

This has been a particularly long and exhausting chase; and there is no crock of gold awaiting us this time either. But, finally looking back now at this paper rainbow, I still love it, for it is my rainbow, and I painted it myself.

> To conclude, I think that there is only one way to science—or to philosophy, for that matter: to meet a problem, to see its beauty and fall in love with it; to get married to it, and to live with it happily, till death do ye part—unless you should meet another and even more fascinating problem, or unless, indeed, you should obtain a solution. But even if you do obtain a solution, you may then discover, to your delight, the existence of a whole family of enchanting though perhaps difficult problem children for whose welfare you may work, with a purpose, to the end of your days.
>
> Popper (1983, Preface 1956, p. 8)

Epilogue

"Are you going to listen to what I am telling you about the Leprecaun?" said the Thin Woman.

"I am not," said the Philosopher. "It has been suggested that we go to sleep at night because it is then too dark to do anything else; but owls, who are a venerably sagacious folk, do not sleep in the nighttime. Bats, also, are a very clear-minded race; they sleep in the broadest day, and they do it in a charming manner. They clutch the branch of a tree with their toes and hang head downwards—a position which I consider singularly happy, for the rush of blood to the head consequent on this inverted position should engender a drowsiness and a certain imbicility of mind which must either sleep or explode."

"Will you never be done talking?" shouted the Thin Woman passionately.

"I will not," said the Philosopher. "In certain ways sleep is useful. It is an excellent way of listening to an opera or seeing pictures on a bioscope. As a medium for day-dreams I know of nothing that can equal it. As an accomplishment it is graceful, but as a means of spending a night it is intolerably ridiculous. If you were going to say anything, my love, please say it now, but you should always remember to think before you speak. A woman should be seen seldom but never heard. Quietness is the beginning of virtue. To be silent is to be beautiful. Stars do not make a noise. Children should always be in bed. These are serious truths, which cannot be controverted; therefore, silence is fitting as regards them."

"Your stirabout is on the hob," said the Thin Woman. "You can get it for yourself. I would not move the breadth of my nail if you were dying of hunger. I hope there's lumps in it. A Leprecaun from Gort na Cloca Mora was here to-day. They'll give it to you for robbing their pot of gold. You old thief, you! you lob-eared, crock-kneed fat-eye!"

The Thin Woman whizzed suddenly from where she stood and leaped into bed. From beneath the blanket she turned a vivid, furious eye on her husband. She was trying to give him rheumatism and toothache and lockjaw all at once. If she had been satisfied to concentrate all her attention on one only of these torments she might have succeeded in afflicting her husband according to her wish, but she was not able to do that.

"Finality is death. Perfection is finality. Nothing is perfect. There are lumps in it," said the Philosopher.

James Stephens
The Crock of Gold

Bibliography

Arbib, Michael A. 1969a. Self-Reproducing Automata—Some Implications for Theoretical Biology. *Pages 204–226 of:* Waddington, C. H. (ed), *Towards a Theoretical Biology, 2: Sketches.* Edinburgh: Edinburgh University Press.

Arbib, Michael A. 1969b. *Theories of Abstract Automata.* Prentice-Hall Series in Automatic Computation. Englewood Cliffs, New Jersey: Prentice-Hall Inc.

Asimov, Issac. 1976. *The Bicentennial Man.* London: Granada Publishing. First published in Great Britain by Victor Gollancz Ltd 1977.

Ayala, Francisco Jose, & Dobzhansky, Theodosius (eds). 1974. *Studies in the Philosophy of Biology.* London: The Macmillan Press Ltd.

Bartley, III, W. W. 1987. A Refutation of the Alleged Refutation of Comprehensively Critical Rationalism. *Chap. XV, pages 313–341 of:* (Radnitzky & Bartley 1987).

Beer, Randall D. 1990. *Intelligence as Adaptive Behavior: An Experiment in Computational Neuroethology.* Series: Perspectives in Artificial Intelligence, vol. 6. Boston: Academic Press, Inc. Series Editor: B. Chandrasekaran.

Beer, Stafford. 1973. Preface to 'Autopoiesis: The Organization of the Living'. *Pages 63–72 of:* (Maturana & Varela 1973).

Beloff, J. 1965. The Identity Hypothesis: A Critique. *In:* Smythies, J. R. (ed), *Brain and Mind.* London: Routledge & Kegan Paul. As cited by Popper & Eccles (1977).

Bendall, D. S. (ed). 1983. *Evolution from Molecules to Men.* Cambridge: Cambridge University Press.

Berlekamp, E. R., Conway, J. H., & Guy, R. K. 1982. What is Life? *Chap. 25, pages 817–850 of: Winning Ways for your Mathematical Plays*, vol. 2. London: Academic Press.

Boden, Margaret A. 1984. Failure is not the Spur. *Pages 305–315 of:* Selfridge, Oliver G., Rissland, Edwina L., & Arbib, Michael A. (eds), *Adaptive Control of Ill-Defined Systems.* New York: Plenum Press. Proceedings of the NATO Advanced Research Institute on Adaptive Control of Ill Defined Systems, held June 21–26, 1981, in Moretonhampstead, Devon, England.

Boden, Margaret A. 1988. *Computer Models of Mind.* Cambridge: Cambridge University Press.

Boden, Margaret A. (ed). 1990. *The Philosophy of Artificial Intelligence.* Oxford Readings in Philosophy. Oxford: Oxford University Press.

Brooks, Rodney A. 1986 (May). *Achieving Artificial Intelligence Through Building Robots.* A.I. Memo 899. Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Burks, Arthur W. 1960. Computation, Behavior and Structure in Fixed and Growing Automata. *Pages 282–311 of:* (Yovits & Cameron 1960).

Burks, Arthur W. 1966a. Automata Self-Reproduction. *Pages 251–296 of:* (Burks 1966d).

Burks, Arthur W. 1966b. Editor's Introduction to 'Theory of Self-Reproducing Automata'. *Pages 1–28 of:* (Burks 1966d).

Burks, Arthur W. 1966c. Preface to 'Theory of Self-Reproducing Automata'. *Pages xv–xix of:* (Burks 1966d).

Burks, Arthur W. (ed). 1966d. *Theory of Self-Reproducing Automata [by] John von Neumann.* Urbana: University of Illinois Press.

Burks, Arthur W. (ed). 1970a. *Essays on Cellular Automata.* Urbana: University of Illinois Press.

Burks, Arthur W. 1970b. Introduction to 'Essays on Cellular Automata'. *Pages xi–xxvi of:* (Burks 1970a).

Burks, Arthur W. 1970c. Toward a Theory of Automata Based on More Realistic Primitive Elements. *Pages 84–102 (Essay Three) of:* (Burks 1970a).

Burks, Arthur W. 1970d. Von Neumann's Self-Reproducing Automata. *Pages 3–64 (Essay One) of:* (Burks 1970a).

Cairns-Smith, A. G. 1982. *Genetic Takeover and the Mineral Origins of Life.* Cambridge: Cambridge University Press.

Campbell, Bill, & Fejer, Peter. 1991. The 1991 Loebner Prize Competition. *IEEE Expert*, **6**(6), 47–48.

Campbell, Donald T. 1960a. Blind Variation and Selective Retention in Creative Thought as in Other Knowledge Processes. *Psychological Review*, **67**(6), 380–400. Also reprinted as *Chap. III, pages 91–114 of:* (Radnitzky & Bartley 1987).

Campbell, Donald T. 1960b. Blind Variation and Selective Survival as a General Strategy in Knowledge-Processes. *Pages 205–231 of:* (Yovits & Cameron 1960).

Campbell, Donald T. 1974a. Evolutionary Epistemology. *Pages 413–463 (Book I) of:* (Schilpp 1974). Also reprinted as *Chap. II, pages 47–89 of:* (Radnitzky & Bartley 1987).

Campbell, Donald T. 1974b. Unjustified Variation and Selective Retention in Scientific Discovery. *Chap. 9, pages 139–161 of:* (Ayala & Dobzhansky 1974).

Charniak, Eugene, & McDermott, Drew. 1985. *Introduction to Artificial Intelligence.* Reading, Massachusetts: Addison-Wesley. World Student Series Edition.

Churchland, Paul M., & Churchland, Patricia Smith. 1990. Could a Machine Think? *Scientific American*, **262**(1), 26–31.

Cliff, D. T. 1990 (May). *Computational Neuroethology: A Provisional Manifesto.* Cognitive Science Research Paper CSRP 162. University of Sussex, School of Cognitive and Computing Sciences.

Codd, E. F. 1968. *Cellular Automata.* ACM Monograph Series. New York: Academic Press, Inc.

Davis, Martin (ed). 1965. *The Undecidable.* New York: Raven Press.

Dawkins, Richard. 1976. *The Selfish Gene.* Oxford: Oxford University Press. See also Dawkins (1989b).

Dawkins, Richard. 1983. Universal Darwinism. *Chap. 20, pages 403–425 of:* (Bendall 1983).

Dawkins, Richard. 1986. *The Blind Watchmaker.* London: Penguin Books.

Dawkins, Richard. 1989a. The Evolution of Evolvability. *Pages 201–220 of:* (Langton 1989a).

Dawkins, Richard. 1989b. *The Selfish Gene.* New edn. Oxford: Oxford University Press. This is a significantly revised edition of (Dawkins 1976), with endnotes and two new chapters added.

Dennett, Daniel C. 1970. The Abilities of Men and Machines. *Chap. 13, pages 256–266 of:* (Dennett 1978b). First presented December 1970 at the American Philosophical Association Eastern Division Meeting.

Dennett, Daniel C. 1971. Intentional Systems. *Chap. 1, pages 3–22 of:* (Dennett 1978b). First published 1971, *Journal of Philosophy*, **LXVIII**(4) 87–106.

Dennett, Daniel C. 1973. Mechanism and Responsibility. *Chap. 12, pages 234–255 of:* (Dennett 1978b). First published 1973, in Honderich, Ted (ed), *Essays on Freedom of Action.* Routledge & Kegan Paul.

Dennett, Daniel C. 1975. Why the Law of Effect Will Not Go Away. *Chap. 5, pages 71–89 of:* (Dennett 1978b). First published 1975, *Journal of the Theory of Social Behaviour*, **V**(2) 169–187.

Dennett, Daniel C. 1976. Where Am I? *Chap. 17, pages 310–323 of:* (Dennett 1978b). First presented October 1976 at the Chapel Hill Coloquium.

Dennett, Daniel C. 1977a. A Cure for the Common Code. *Chap. 6, pages 90–108 of:* (Dennett 1978b). First published as "Critical Notice: *The Language of Thought* by Jerry Fodor", *Mind*, April 1977.

Dennett, Daniel C. 1977b. Two Approaches to Mental Images. *Chap. 10, pages 174–189 of:* (Dennett 1978b). First presented October 1977 at the Western Canadian Philosophical Association at the University of Manitoba, Winnipeg.

Dennett, Daniel C. 1978a. Artificial Intelligence as Philosophy and as Psychology. *Chap. 7, pages 109–126 of:* (Dennett 1978b). First published 1978, in Ringle, Martin (ed), *Philosophical Perspectives on Artificial Intelligence.* New York: Humanities Press and Harvester Press.

Dennett, Daniel C. 1978b. *Brainstorms: Philosophical Essays on Mind and Psychology.* Brighton, Sussex: The Harvester Press Limited. Harvester Press edition first published in 1981.

Dennett, Daniel C. 1978c. Toward a Cognitive Theory of Consciousness. *Chap. 9, pages 149–173 of:* (Dennett 1978b). First published 1978, in Wade Savage, C. (ed), *Perception and Cognition: Issues in the Foundations of Psychology, Minnesota Studies in the Philosophy of Science* IX. University of Minnesota Press.

Dennett, Daniel C. 1978d. Why Not the Whole Iguana? *The Behavioral and Brain Sciences*, **1**, 103–104. Commentary on Pylyshyn: 'Computational Models and Empirical Constraints'.

Dennett, Daniel C. 1979. Book Review: The Self and Its Brain. *The Journal of Philosophy*, **76**(2), 91–97. Review of (Popper & Eccles 1977).

Dennett, Daniel C. 1980. The Milk of Human Intentionality. *The Behavioral and Brain Sciences*, **3**, 428–430. Commentary on (Searle 1980).

Dennett, Daniel C. 1981. Three Kinds of Intentional Psychology. *Chap. 3, pages 43–68 of:* (Dennett 1987). First published 1981, in Healy, R. (ed), *Reduction, Time and Reality.* Cambridge: Cambridge University Press.

Dennett, Daniel C. 1984. *Elbow Room: The Varieties of Free Will Worth Wanting.* Oxford: Clarendon Press.

Dennett, Daniel C. 1986. *Content and Consciousness.* Second edn. International Library of Philosophy and Scientific Method. London: Routledge & Kegan Paul plc. First edition published 1969.

Dennett, Daniel C. 1987. *The Intentional Stance.* Cambridge, Massachusetts: The MIT Press.

Dewdney, A. K. 1987. Computer Recreations: A Program Called MICE Nibbles its Way to Victory at the First Core Wars Tournament. *Scientific American*, **256**(1), 8–11.

Dodd, T. 1991. Gödel, Penrose and the possibility of AI. *Artificial Intelligence Review*, **5**, 187–199.

Dreyfus, Hubert L., & Dreyfus, Stuart E. 1986. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer.* Oxford: Basil Blackwell Ltd.

Dyson, Freeman. 1985. *Origins Of Life.* Cambridge: Cambridge University Press.

Eccles, John C. 1980. A Dualist-Interactionist Perspective. *The Behavioral and Brain Sciences*, **3**, 430–431. Commentary on (Searle 1980).

Eigen, Manfred, & Schuster, Peter. 1979. *The Hypercycle: A Principle of Natural Self-Organization.* Berlin: Springer-Verlag.

Farmer, J. Doyne, & d'A. Belin, Alleta. 1992. Artificial Life: The Coming Evolution. *Pages 815–838 of:* (Langton *et al.* 1992).

Fodor, Jerry A. 1976. *The Language of Thought.* Hassocks, Sussex: The Harvester Press Limited. First published in the United States of America by Thomas Y. Crowell Company, Inc.

Fodor, Jerry A., & Pylyshyn, Zenon W. 1988. Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition*, **28**, 3–71.

Fogel, Lawrence J., Owens, Alvin J., & Walsh, Michael J. 1966. *Artificial Intelligence Through Simulated Evolution.* New York: John Wiley & Sons, Inc.

French, Robert M. 1990. Subcognition and the Limits of the Turing Test. *Mind*, **99**(393), 53–65.

Friedberg, R. M. 1958. A Learning Machine: Part I. *IBM Journal*, Jan., 2–13.

Friedberg, R. M., Dunham, B., & North, J. H. 1959. A Learning Machine: Part II. *IBM Journal*, July, 282–287.

Goldberg, David E. 1989. *Genetic Algorithms in Search, Optimization & Machine Learning.* Reading, Massachusetts: Addison-Wesley Publishing Company, Inc.

Gould, Stephen Jay. 1982. Darwinism and the Expansion of Evolutionary Theory. *Science*, **216**(23 April), 380–387.

Harnad, Stevan. 1989. Minds, Machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence*, **1**, 5–25.

Harnad, Stevan. 1990. The Symbol Grounding Problem. *Physica*, **42D**, 335–346.

Haugeland, John (ed). 1981. *Mind Design.* Cambridge: MIT Press.

Hayes, Patrick J. 1979. The Naïve Physics Manifesto. *Chap. 8, pages 171–205 of:* (Boden 1990). First published 1979 as *pages 242–270 of:* Mitchie, D. (ed), *Expert Systems in the Micro-Electronic Age.* Edinburgh: Edinburgh University Press.

Herman, Gabor T. 1973. On Universal Computer-Constructors. *Information Processing Letters*, **2**, 61–64.

Hodges, Andrew. 1983. *Alan Turing: The Enigma of Intelligence.* London: Unwin Paperbacks. First published in Great Britain by Burnett Books Limited in association with the Hutchinson Publishing Group 1983. First published in Unwin Paperbacks 1985.

Hofstadter, Douglas R. 1979. *Gödel, Escher, Bach: An Eternal Golden Braid.* London: Penguin Books. First published in Great Britain by The Harvester Press Ltd 1979. Published in Penguin Books 1980.

Hofstadter, Douglas R. 1983. Artificial Intelligence: Subcognition as Computation. *Pages 263–285 of:* (Machlup & Mansfield 1983). Also reprinted, with a new *Post Scriptum*, as *Chap. 26, pages 631–665 of:* (Hofstadter 1985).

Hofstadter, Douglas R. 1985. *Metamagical Themas: Questing for the Essence of Mind and Pattern.* London: Penguin Books. Published in Penguin Books 1986.

Hofstadter, Douglas R., & Dennett, Daniel C. (eds). 1981. *The Mind's I: Fantasies and Reflections on Self and Soul.* Harmondsworth, Middlesex: Penguin Books Ltd. Published in Penguin Books 1982.

Holland, John H. 1962a. Concerning Efficient Adaptive Systems. *Pages 215–230 of:* Yovits, Marshall C., Jacobi, George T., & Goldstein, Gordon D. (eds), *Self-Organizing Systems 1962.* Washington D.C.: Spartan Books.

Holland, John H. 1962b. Outline for a Logical Theory of Adaptive Systems. *Journal of the ACM*, **9**(3), 297–314.

Holland, John H. 1975. *Adaptation in Natural and Artificial Systems.* Ann Arbor: The University of Michigan Press.

Holland, John H. 1976. Studies of the Spontaneous Emergence of Self-Replicating Systems Using Cellular Automata and Formal Grammars. *Pages 385–404 of:* Lindenmayer, A., & Rozenberg, G. (eds), *Automata, Languages, Development.* New York: North-Holland.

Holland, John H. 1986. Escaping Brittleness: The Possibilities of General-Purpose Learning Algorithms Applied to Parallel Rule-Based Systems. *Chap. 20, pages 593–623 of:* Michalski, Ryszard S., Carbonell, Jaime G., & Mitchell, Tom M. (eds), *Machine Learning: An Artificial Intelligence Approach: Volume II.* Los Altos, California: Morgan Kaufman Publishers Inc.

Holland, John H., Holyoak, Keith J., Nisbett, Richard E., & Thagard, Paul R. 1986. *Induction.* Series: Computational Models of Cognition and Perception. Cambridge: The MIT Press. Series Editors: Jerome A. Feldman, Patrick J. Hayes, and David E. Rumelhart. First MIT Press paperback edition, 1989.

Huxley, Andrew. 1983. How Far Will Darwin Take Us? *Chap. 1, pages 3–19 of:* (Bendall 1983).

Kampis, G., & Csányi, V. 1987. Replication in Abstract and Natural Systems. *BioSystems*, **20**, 143–152.

Kampis, George. 1991. *Self-Modifying Systems in Biology and Cognitive Science.* IFSR International Series on Systems Science and Engineering, vol. 6. Oxford: Pergamon Press. Editor-in-Chief: George J. Klir.

Kauffman, Stuart A. 1990. Requirements for Evolvability in Complex Systems: Orderly Dynamics and Frozen Components. *Physica*, **42D**, 135–152.

Kemeny, John G. 1955. Man Viewed as a Machine. *Scientific American*, **192**(4), 58–67.

Koestler, Arthur. 1959. *The Sleepwalkers.* Harmondsworth, Middlesex: Penguin Books Ltd. First Published 1959, by Hutchinson. Published in Penguin Books 1964. Reissued in Pelican Books 1968.

Laing, Richard A. 1975 (Aug.). Artificial Molecular Machines: A Rapprochement Between Kinematic and Tessellation Automata. *Pages 73–80 of: Proceedings of the International Symposium on Uniformly Structured Automata and Logic, Tokyo.*

Laing, Richard A. 1977. Automaton Models of Reproduction by Self-Inspection. *Journal of Theoretical Biology*, **66**, 437–456.

Langton, Christopher G. 1984. Self-Reproduction in Cellular Automata. *Physica*, **10D**, 135–144.

Langton, Christopher G. 1986. Studying Artificial Life with Cellular Automata. *Physica*, **22D**, 120–149.

Langton, Christopher G. (ed). 1989a. *Artifical Life.* Series: Sante Fe Institute Studies in the Sciences of Complexity, vol. VI. Redwood City, California: Addison-Wesley Publishing Company, Inc. Proceedings of an interdisciplinary workshop on the synthesis and simulation of living systems held September, 1987, in Los Alamos, New Mexico.

Langton, Christopher G. 1989b. Artificial Life. *Pages 1–47 of:* (Langton 1989a).

Langton, Christopher G., Taylor, Charles, Farmer, J. Doyne, & Rasmussen, Steen (eds). 1992. *Artifical Life II.* Series: Sante Fe Institute Studies in the Sciences of Complexity, vol. X. Redwood City, California: Addison-Wesley Publishing Company, Inc. Proceedings of the workshop on Artificial Life held February, 1990, in Sante Fe, New Mexico.

Larson, Harold J. 1974. *Introduction to Probability Theory and Statistical Inference.* Wiley Series in Probability and Mathematical Statistics. New York: John Wiley & Sons.

Lenat, Douglas B. 1983. The Role of Heuristics in Learning by Discovery: Three Case Studies. *Chap. 9, pages 243–306 of:* Michalski, Ryszard S., Carbonell, Jaime G., & Mitchell, Tom M. (eds), *Machine Learning: An Artificial Intelligence Approach.* Palo Alto, California: Tioga Publishing Company.

Lenat, Douglas B., & Guha, R. V. 1990. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project.* Reading, Massachusetts: Addison-Wesley Publishing Company, Inc.

Lewis, Harry R., & Papadimitriou, Christos H. 1981. *Elements of the Theory of Computation.* Prentice-Hall Software Series. London: Prentice-Hall International. Brian W. Kernighan, advisor.

Lindsay, Robert K. 1968. Artificial Evolution of Intelligence. *Contemporary Psychology*, **13**(3), 113–116. Review of (Fogel *et al.* 1966).

Lucas, J. R. 1961. Minds, Machines and Gödel. *Philosophy*, **36**, 112–127.

Machlup, Fritz, & Mansfield, Una (eds). 1983. *The Study of Information.* New York: John Wiley & Sons.

Magee, Bryan. 1973. *Popper.* London: Fontana Paperbacks.

Martinez, Hugo M. 1979. An Automaton Analog of Unicellularity. *BioSystems*, **11**, 133–162.

Maturana, Humberto R., & Varela, Francisco J. 1973. Autopoiesis: The Organization of the Living. *Pages 59–138 of:* (Maturana & Varela 1980). Dated 1973. First published 1972 in Chile under the title *De Maquinas y Seres Vivos*, Editorial Universitaria S.A.

Maturana, Humberto R., & Varela, Francisco J. 1980. *Autopoiesis and Cognition.* Series: Boston Studies in the Philosophy of Science, vol. 42. Dordrecht, Holland: D. Reidel Publishing Company. With a preface to 'Autopoiesis' by Stafford Beer. Series editors: Robert S. Cohen and Marx W. Wartofsky.

Maynard Smith, John. 1970. Natural Selection and the Concept of a Protein Space. *Nature*, **225**(Febuary 7), 563–564.

McDermott, Drew. 1976. Artificial Intelligence Meets Natural Stupidity. *Chap. 5, pages 143–160 of:* (Haugeland 1981). First published 1976: *SIGART Newsletter*, No. 57 (April).

McDermott, Drew. 1987. A Critique of Pure Reason. *Chap. 9, pages 206–230 of:* (Boden 1990). First published 1987, *Computational Intelligence,* **3** 151–160.

McMullin, Barry. 1988 (Sept.). *Darwinism Applied to Machine Learning.* Technical Report `NIHED/EE/88-11`. School of Electronic Engineering, Dublin City University (then the National Institute for Higher Education, Dublin), Dublin 9, Ireland.

McMullin, Barry. 1989 (May). *Computational Darwinism: A Research Proposal.* Technical Report `NIHED/EE/89-11`. School of Electronic Engineering, Dublin City University (then the National Institute for Higher Education, Dublin), Dublin 9, Ireland.

McMullin, Barry. 1990. Computational Darwinism, or Who Teaches the Teacher? *Pages 211–231 of:* Smeaton, Alan F., & McDermott, Gabriel (eds), *AI and Cognitive Science '89.* London: Springer-Verlag.

McMullin, Barry. 1992a (Mar.). *Essays on Darwinism. 1: Ontological Foundations.* Technical Report `bmcm9201`. School of Electronic Engineering, Dublin City University, Dublin 9, Ireland.

McMullin, Barry. 1992b (Apr.). *Essays on Darwinism. 2: Organismic Darwinism.* Technical Report `bmcm9202`. School of Electronic Engineering, Dublin City University, Dublin 9, Ireland.

McMullin, Barry. 1992c (May). *Essays on Darwinism. 3: Genic and Organismic Selection.* Technical Report `bmcm9203`. School of Electronic Engineering, Dublin City University, Dublin 9, Ireland.

McMullin, Barry. 1992d. The Holland α-Universes Revisited. *Pages 317–326 of:* (Varela & Bourgine 1992).

Minsky, Marvin. 1986. *The Society of Mind.* London: Pan Books Ltd. First published in Great Britain 1987 by William Heinemann Ltd. Picador edition published 1988.

Minsky, Marvin L. 1967. *Computation: Finite and Infinite Machines.* Prentice-Hall Series in Automatic Computation. Englewood Cliffs, New Jersey: Prentice-Hall Inc.

Montague, R. 1974. *Formal Philosophy: Selected Papers.* New Haven: Yale University Press. As cited by (Boden 1988).

Moore, Edward F. 1970. Machine Models of Self-Reproduction. *Pages 187–203 (Essay Six) of:* (Burks 1970a).

Moorhead, Paul S., & Kaplan, Martin M. (eds). 1967. *Mathematical Challenges to the Neo-Darwinian Interpretation of Evolution.* Philadelphia: The Wistar Institute Press, for The Wistar Institute of Anatomy and Biology.

Mühlenbein, H. 1989. The Dynamics of Evolution and Learning—Towards Genetic Neural Networks. *In:* Pfeifer, Rolf, Schreter, Z., Fogelman-Soulie, F., & Steels, L. (eds), *Connectionism in Perspective.* North-Holland.

Mühlenbein, H., Gorges-Schleuter, M., & Krämer, O. 1988. Evolution Algorithms in Combinatorial Optimization. *Parallel Computing*, **7**, 65–85.

Mühlenbein, Heinz. 1992. Darwin's Continent Cycle Theory and its Simulation by the Iterated Prisoner's Dilemma. *Pages 236–244 of:* (Varela & Bourgine 1992).

Newell, Allen. 1983. Endnotes to the Papers on Artificial Intelligence. *Pages 287–294 of:* (Machlup & Mansfield 1983).

Newell, Allen, & Simon, Herbert A. 1972. *Human Problem Solving.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Newell, Allen, & Simon, Herbert A. 1976. Computer Science as Empirical Enquiry: Symbols and Search. *Chap. 1, pages 35–66 of:* (Haugeland 1981). First published 1976: *Communications of the ACM*, **19** (March) 113–126.

Newton-Smith, W. 1981. In Defence of Truth. *Chap. 10, pages 269–289 of:* Jensen, U. J., & Harré, R. (eds), *The Philosophy of Evolution.* Harvester Studies in Philosophy. Brighton, Sussex: The Harvester Press Limited. General Editor: Margaret A. Boden.

Oparin, A. I. 1953. *Origin Of Life.* New York: Dover Publications, Inc.

Packard, Norman H. 1989. Intrinsic Adaptation in a Simple Model for Evolution. *Pages 141–155 of:* (Langton 1989a).

Park, Stephen K., & Miller, Keith W. 1988. Random Number Generators: Good Ones are Hard to Find. *Communications of the ACM*, **31**(10), 1192–1201.

Penrose, Roger. 1990. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics.* London: Vintage.

Popper, Karl R. 1949. The Bucket and the Searchlight: Two Theories of Knowledge. *Pages 153–190 (Appendix 1) of:* (Popper 1979). First published 1949 (in German) as "Naturgesetze und theoretische Systeme" in Moser, Simon (ed), *Gesetz und Wirklichkeit.*

Popper, Karl R. 1953. Language and the Body-Mind Problem. *Chap. 12, pages 293–298 of:* (Popper 1989). First published 1953 in the *Proceedings of the 11th International Congress of Philosophy* **7**.

Popper, Karl R. 1961. Evolution and the Tree of Knowledge. *Chap. 7, pages 256–280 of:* (Popper 1979). Based on the Herbert Spencer Lecture, delivered in Oxford on 30 Oct. 1961.

Popper, Karl R. 1965. Of Clouds and Clocks. *Chap. 6, pages 206–255 of:* (Popper 1979). This was the second Arthur Holly Compton Memorial Lecture, presented at Washington University on 21 Apr. 1965.

Popper, Karl R. 1968. Epistemology Without a Knowing Subject. *Chap. 3, pages 106–152 of:* (Popper 1979). First published 1968 as *pages 333–373 of:* van Rootselaar, B., and Staal, J.F. (eds), *Proceedings of the Third International Congress for Logic, Methodology and Philosophy of Science.* 25 Aug. to 2 Sept. 1967. Amsterdam.

Popper, Karl R. 1970a. On the Theory of the Objective Mind. *Chap. 4, pages 153–190 of:* (Popper 1979). Reproduced from the *Akten des XIV. Internationalen Kongresses für Philosophie,* vol. i, Vienna 1968, pp. 25–53. Includes additional material first published (in German) in *Schweizer Monatshefte,* 50. Jahr, Heft 3, 1970, pp. 207–215.

Popper, Karl R. 1970b. Two Faces of Common Sense: An Argument for Commonsense Realism and Against the Commonsense Theory of Knowledge. *Chap. 2, pages 32–105 of:* (Popper 1979). This is a revised and expanded version of a talk first given by Popper in 1970, to his former Seminar.

Popper, Karl R. 1971. Conjectural Knowledge: My Solution of the Problem of Induction. *Chap. 1, pages 1–31 of:* (Popper 1979). First published 1971 *Revue internationale de Philosophie*, 25^{e} année, no. 95–6, fasc. 1–2.

Popper, Karl R. 1973. Indeterminism is not Enough. *Encounter*, **40**(4), 20–26. A revised version of this essay appears as *Addendum 1, pages 113–130, of:* Popper (1988).

Popper, Karl R. 1974a. Autobiography of Karl Popper. *Pages 1–181 (Book I) of:* (Schilpp 1974). See also (Popper 1976).

Popper, Karl R. 1974b. Replies to my Critics. *Pages 961–1197 (Book II) of:* (Schilpp 1974).

Popper, Karl R. 1974c. Scientific Reduction and the Essential Incompleteness of All Science. *Chap. 16, pages 259–284 of:* (Ayala & Dobzhansky 1974). A revised version of this essay appears as *Addendum 2, pages 131–162, of:* Popper (1988).

Popper, Karl R. 1976. *Unended Quest.* Glasgow: Fontana/William Collins Sons & Co. Ltd. First edition published as (Popper 1974a). This revised edition first published 1976 by Fontana.

Popper, Karl R. 1979. *Objective Knowledge: An Evolutionary Approach.* Oxford: Oxford University Press. Revised edition (reprinted with corrections and a new appendix 2). First edition published 1972.

Popper, Karl R. 1980. *The Logic of Scientific Discovery.* Tenth impression (revised) edn. London: Unwin Hyman Ltd. First edition published 1959 by Hutchinson Education. Translation of *Logik der Forschung*, published in Vienna in the autumn of 1934.

Popper, Karl R. 1983. *Realism and the Aim of Science.* London: Hutchinson. From the *Postscript to the Logic of Scientific Discovery*, edited by W.W. Bartley, III. Note that the original text dates largely from the period 1951–1956.

Popper, Karl R. 1988. *The Open Universe: An Argument for Indeterminism.* London: Hutchinson. From the *Postscript to the Logic of Scientific Discovery*, edited by W.W. Bartley, III. First edition published 1982. Note that the original text dates largely from the period 1951–1956.

Popper, Karl R. 1989. *Conjectures and Refutations.* Fifth (revised) edn. London: Routledge & Kegan Paul. First edition published 1963.

Popper, Karl R., & Eccles, John C. 1977. *The Self and its Brain: An Argument for Interactionism.* London: Routledge & Kegan Paul plc. First published 1977, Berlin: Springer-Verlag. This edition first published 1983.

Radnitzky, Gerard, & Bartley, III, W. W. (eds). 1987. *Evolutionary Epistemology, Rationality, and the Sociology of Knowledge.* La Salle, Illinois: Open Court.

Rasmussen, Steen, Knudsen, Carsten, Feldberg, Rasmus, & Hindsholm, Morten. 1990. The Coreworld: Emergence and Evolution of Cooperative Structures in a Computational Chemistry. *Physica*, **42D**, 111–134.

Ray, Thomas S. 1991. Tierra Update. *Alife Digest*, November 5th. *Alife Digest* is an electronic newsletter distributed over `Internet`; for information contact `alife-request@cognet.ucla.edu`.

Ray, Thomas S. 1992. An Approach to the Synthesis of Life. *Pages 371–408 of:* (Langton *et al.* 1992).

Reeke, Jr., George N., & Edelman, Gerald M. 1988. Real Brains and Artificial Intelligence. *Proceedings of the American Academy of Arts and Sciences*, **117**(1), 143–173. Daedalus, Winter 1988.

Regis, Ed. 1987. *Who Got Einstein's Office? Eccentricity and Genius at the Institute for Advanced Study.* London: Penguin Group. First published in the USA by Addison-Wesley Co, Inc. 1987. First published in Great Britain by Simon & Schuster 1988. Published in Penguin Books 1989.

Rizki, Mateen M., & Conrad, Michael. 1985. Evolve III: A Discrete Events Model of an Evolutionary Ecosystem. *BioSystems*, **18**, 121–133.

Rosen, Robert. 1959. On a Logical Paradox Implicit in the Notion of a Self-Reproducing Automaton. *Bulletin of Mathematical Biophysics*, **21**, 387–394.

Rosen, Robert. 1985a. *Anticipatory Systems.* IFSR International Series on Systems Science and Engineering, vol. 1. Oxford: Pergamon Press. Editor-in-Chief: George J. Klir.

Rosen, Robert. 1985b. Organisms as Causal Systems Which Are Not Mechanisms: An Essay into the Nature of Complexity. *Chap. 3, pages 165–203 of:* Rosen, Robert (ed), *Theoretical Biology and Complexity.* Orlando: Academic Press, Inc.

Schaffer, J. David, & Greffenstette, John J. 1988. A Critical Review of Genetic Algorithms. *In: Critical Reviews in Artificial Intelligence.* Cleveland: CRC Press.

Schilpp, Paul Arthur (ed). 1974. *The Philosophy of Karl Popper.* The Library of Living Philosophers, vol. XIV. Illinois: Open Court.

Schwefel, Hans-Paul. 1979. Direct Search for Optimal Parameters within Simulation Models. *Pages 91–102 of: Proceedings of the 12th Annual Simulation Symposium.* (Tampa, Florida, March 14–16, 1979).

Schwefel, Hans-Paul. 1988. Evolutionary Learning Optimum-Seeking on Parallel Computer Architectures. *Pages 217–225 of: Systems Analysis and Simulation 1988. I: Theory and Foundations.* Berlin: Akademie-Verlag. Proceedings of the International Symposium held in Berlin (GDR), September 12–16, 1988.

Searle, John R. 1980. Minds, Brains, and Programs. *The Behavioral and Brain Sciences*, **3**, 417–457. Includes peer commentaries. Also reprinted (without commentaries) as *Chap. 10, pages 282–305 of:* (Haugeland 1981) and as *Chap. 3, pages 67–88 of:* (Boden 1990).

Searle, John R. 1990. Is the Brain's Mind a Computer Program? *Scientific American*, **262**(1), 20–25.

Selfridge, O. G. 1959. Pandemonium: A Paradigm for Learning. *Pages 511–531 of: Mechanisation of Thought Processes.* London: HMSO.

Simon, Herbert A. 1969. *The Sciences of the Artificial.* Cambridge: The MIT Press. The Karl Taylor Compton Lectures, MIT, 1968. See also (Simon 1981).

Simon, Herbert A. 1981. *The Sciences of the Artificial.* Second edn. Cambridge: The MIT Press. This is a revised and expanded edition of (Simon 1969), now also including the H. Rowan Gaither lectures, University of California, Berkeley, 1980.

Smolensky, Paul. 1991. Connectionism, Constituency, and the Language of Thought. *Chap. 12, pages 201–227 of:* Loewer, Barry, & Rey, Georges (eds), *Meaning in Mind: Fodor and his Critics.* Cambridge, Massachusetts: Basil Blackwell Inc.

Sperry, Roger W. 1965. Mind, Brain, and Humanist Values. *Chap. 4, pages 71–92 of:* Platt, John R. (ed), *New Views on the Nature of Man.* Chicago: The University of Chicago Press.

Strok, Dale. 1991. Computers Try to Fool Human Judges. *IEEE Expert*, **6**(6), 47.

Taub, A. H. (ed). 1961. *John von Neumann: Collected Works. Volume V: Design of Computers, Theory of Automata and Numerical Analysis.* Oxford: Pergamon Press.

Taylor, Charles E., Jefferson, David R., Turner, Scott R., & Goldman, Seth R. 1989. RAM: Artificial Life for the Exploration of Complex Biological Systems. *Pages 275–295 of:* (Langton 1989a).

Thatcher, J. W. 1970. Universality in the von Neumann Cellular Model. *Pages 132–186 (Essay Five) of:* (Burks 1970a).

Tipler, Frank J. 1981. Extraterrestrial Intelligent Beings Do Not Exist. *Physics Today*, **32**(4), 9, 70–71.

Tipler, Frank J. 1982. We are Alone in our Galaxy. *New Scientist*, **96**(1326), 33–35.

Turing, Alan. 1936. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society*, **Series 2, Vol. 42**, 230–265. Also reprinted, including corrections, as *pages 116–154 of:* (Davis 1965).

Turing, Alan M. 1950. Computing Machinery and Intelligence. *Mind*, **LIX**(236), 433–460. Also reprinted as *Chap. 2, pages 40–66 of:* (Boden 1990).

Varela, Francisco J. 1979. *Principles of Biological Autonomy.* New York: North-Holland.

Varela, Francisco J., & Bourgine, Paul (eds). 1992. *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life.* Series: Complex Adaptive Systems. Cambridge: MIT Press. Series Advisors: John H. Holland, Christopher Langton and Stewart W. Wilson.

Varela, Francisco J., Maturana, Humberto R., & Uribe, R. 1974. Autopoiesis: The Organization of Living Systems, its Characterization and a Model. *BioSystems*, **5**, 187–196.

von Neumann, John. 1951. The General and Logical Theory of Automata. *Chap. 9, pages 288–328 of:* (Taub 1961). First published 1951 as *pages 1–41 of:* L. Jeffress, A. (ed), *Cerebral Mechanisms in Behavior—The Hixon Symposium*, New York: John Wiley.

von Neumann, John. 1966a. Theory and Organization of Complicated Automata. *Pages 29–87 (Part One) of:* (Burks 1966d). Based on transcripts of lectures delivered at the University of Illinois, in December 1949. Edited for publication by A.W. Burks.

von Neumann, John. 1966b. The Theory of Automata: Construction, Reproduction, Homogeneity. *Pages 89–250 of:* (Burks 1966d). Based on an unfinished manuscript by von Neumann. Edited for publication by A.W. Burks.

Weizenbaum, Joseph. 1984. *Computer Power and Human Reason.* London: Penguin Group. First edition published by W.H. Freeman and Company 1976.

Wills, Christopher. 1991. *The Wisdom of the Genes.* Oxford: Oxford University Press.

Winograd, Terry. 1973. A Procedural Model of Language Understanding. *Pages 152–186 of:* Shank, Roger C., & Colby, Kenneth Mark (eds), *Computer Models of Thought and Language.* San Francisco: W. H. Freeman and Company.

Wright, Sewall. 1982. The Shifting Balance Theory and Macroevolution. *Annual Review of Genetics*, **16**, 1–19.

Yovits, Marshall C., & Cameron, Scott (eds). 1960. *Self-Organizing Systems.* Oxford: Pergamon Press. Proceedings of an interdisciplinary conference, 5 and 6 May, 1959.

Zelany, Milan, & Pierre, Norbert A. 1976. Simulation of Self-Renewing Systems. *Chap. 7, pages 150–165 of:* Jantsch, Erich, & Waddington, Conrad H. (eds), *Evolution and Consciousness: Human Systems in Transition.* Reading, Massachusetts: Addison-Wesley Publishing Company.

Zeleny, Milan. 1977. Self-Organization of Living Systems: A Formal Model of Autopoiesis. *International Journal of General Systems*, **4**, 13–28.