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 characteristically 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.