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)