Towards Implementing Free-Will

Bruce Edmonds
Centre for Policy Modelling,
Manchester Metropolitan University,
Aytoun Building, Aytoun Street, Manchester, M1 3GH, UK.
http://www.cpm.mmu.ac.uk/~bruce

"Anyone who considers arithmetic methods of producing random digits is, of course, in a state of sin" John von Neuman[1]

Abstract

Some practical criteria for free-will are suggested where free-will is a matter of degree. It is argued that these are more appropriate than some extremely idealised conceptions. Thus although the paper takes lessons from philosophy it avoids idealistic approaches as irrelevant. A mechanism for allowing an agent to meet these criteria is suggested: that of facilitating the gradual emergence of free-will in the brain via an internal evolutionary process. This meets the requirement that not only must the choice of action be free but also choice in the method of choice, and choice in the method of choice of the method of choice etc. This is directly analogous to the emergence of life from non-life. Such an emergence of indeterminism with respect to the conditions of the agent fits well with the 'Machiavellian Intelligence Hypothesis' which posits that our intelligence evolved (at least partially) to enable us to deal with social complexity and modelling 'arms races'. There is a clear evolutionary advantage in being internally coherent in seeking to fulfil ones goals and unpredictable by ones peers. To fully achieve this vision several other aspects of cognition are necessary: open-ended strategy development; the meta-evolution of the evolutionary process; the facility to anticipate the results of strategies; and the situating of this process in a society of competitive peers. Finally the requirement that reports of the deliberations that lead to actions need to be socially acceptable leads to the suggestion that the language that the strategies are developed within be subject to a normative process in parallel with the development of free-will. An appendix outlines a philosophical position in support of my position.

1 Introduction

To paraphrase the von Neuman quote above: *anyone who considers computational methods of implementing free-will is, of course, in a state of sin*. By simply suggesting that free-will could be implemented I will already have offended the intellectual sensibilities of several groups of people: I will have offended "hard" determinists by suggesting that free-will is possible; I will have offended those who think that free-will is a uniquely human characteristic; and I will have offended those who see free-will as something that is simply beyond design. I have some sympathy for the later two groups – at the moment a human being is the only system that clearly exhibits this facility; and, as will be explained, I do think that free-will can not be *directly* designed into an entity.

Despite almost everybody agreeing that it is fundamentally impossible, arithmetic methods of producing random numbers have become, by far, the most widely used method. These methods (used correctly) are efficient and reliable. We rely on their effective randomness in many cryptographic techniques which, in turn, are relied upon in electronic commerce and the like. Maybe it is time to let the evidence take precedence over assumptive theory – if theory disagrees with practical evidence it is the theory that should change. What was assumed to be a state of sin can turn out to be inspired.

In this paper I will outline a practical architecture that, I argue, could result in a computational entity with free-will. I will start by rejecting extremely idealised conceptions of free-will and suggest instead a more practical set of properties. Then, in section 3, I will put forward the central idea of the paper which is to allow free-will to evolve in a brain during its lifetime. The following 4 sections (4, 5, 6 and 7) consider other

[1] Reportedly said by John von Neuman in 1951 at a conference on Monte Carlo methods.

necessary aspects of the architecture: open-ended development; the co-evolution of strategies against competitive peers; the meta-evolution of the evolutionary process itself; and the necessity of being able to anticipate the consequences of candidate actions. Section 8 then looks at some societal aspects that might allow the development of a framework of acceptable rationality within which free-will can operate. I summarise the suggested architecture in section 9 and finally conclude in section 10. For those who feel philosophically short-changed by this paper there is an appendix which briefly outlines my philosophical position.

2 Conceptions of free-will

It is inevitable that in any implementation process one will move from an *idealised* to a *realised* conception of what one implements. Thus here I am not so interested with artificial or idealised conceptions of free-will, determinism, randomness etc. but with more practical concerns. For if a certain conception of free-will makes no *practical* difference then it is irrelevant to a discussion about implementation (and quite possibly to everything else as well). For if it is impossible to tell whether an entity has a certain property and that entity can do all the things without that property as with it, how can it be relevant in practice?

From this practical perspective, free-will is something that a normal adult human has but an newly fertilised human embryo hasn't. It means that an agent is free to choose according to its will, that is to say that sometimes it is its deliberations on how to achieve its goals that determine its actions and not just its circumstances (including past circumstances).

Of course, many aspects of traditional philosophical analyses of free-will are relevant *if* one avoids the pitfalls of extreme idealisation. For example the points listed below come from philosophy, but are formulated with practical concerns in mind:

(A) The process of deliberation leading to a choice of action has to be free in the sense that it is not constrained to a particular "script" - this means that there is also some *choice* in that deliberation, as well as choice in how to make that choice, and choice in how to make the choice in how to make that choice etc.;

(B) In some circumstances, if others with whom the entity is competing are able to effectively predict its actions they may well exploit this in order to constrain its choice to its detriment - thus it can be important that actions are not predictable by others;

(C) In order for an entity's will to be effective it has to be able to perform some processing that tends to result in actions that (as far as it can tell) furthers its goals – in particular it needs to be able to consider the likely consequences of different possible strategies and *choose* amongst them with a view to furthering its goals;

(D) It must be possible that sometimes it might have taken a different action to those actually taken - that is, given *indistinguishable* circumstances, it would not simply repeat past decisions (even if it did not recall them).

(E) In order to have an entity's decisions allowed by a society of peers it is often necessary that it is able to give an account of its reasons for actions that impinge upon that society, reasons that would be deemed acceptably rational - for those that are not reliably rational can pose a danger to the rest and hence may be prevented from taking certain actions.

These are the criteria I will take to guide my thoughts about implementation rather than abstract issues of theoretical determination and the like. They seem to capture the more *important* aspects of free-will – the aspects of free-will that make it worth having (Dennett 1984).

This is a similar approach to that of Aaron Sloman's (1992), except that it focuses more upon a single issue: *how can we develop an agent whose decisions are determined by its deliberations and* ***not*** *completely constrained by its circumstances*. He is right to point out that an entity's decisions can be constrained in different ways and is dependent upon the capabilities and structure of the entity. However the multiplicity of factors does not dissolve the central issue which is concrete and testable; for any entity placed in the same circumstances one can *measure* the extent to which entity acts in the same way[2] and (with humans) collect indirect evidence (by interview) to see the extent to which the actions correlated with the prior deliberations.

3 Evolving free-will in a brain

The basic idea I am proposing, is to provide, in a constructed brain, an environment which is conducive to the *evolution* of free-will in that brain. In this evolutionary process practical indeterminacy emerges first in infinitesimal amounts and then develops into full-blown adult free-will by degrees. This evolution happens in parallel to the development of rationality in

[2] A practical way of putting this is: *as a circumstance approaches a previous circumstance in similarity, does the probability of the a different action resulting decrease to zero?*

the individuality, so that the result is a will which is internally coherent in furthering its goals but yet not determined by its circumstances.

Those who insist that free-will requires prior free-will (arguing that otherwise the choice process would also be determined) can follow the chain of causation (and indeterminism) backwards until it slowly diminishes down a limit of nothing (determinism). In this model the gradual emergence of free-will in the brain is analogous to the emergence of life - it can start from infinitesimal amounts and evolve up from there. This requires that free-will can come in different *degrees* – that circumstances can constrain behaviour to different extents from totally (determinism) to partially (some degree of indetermination). The artificiality of an all-or-nothing division into *having it* or *not* makes as little sense with free-will as it does with life, especially if one is discussing mechanisms for its appearance (as must occur somewhere between the newly fertilised embryo and the adult human. As Douglas Hofstadter said (1985):

> *Perhaps the problem is the seeming need that people have of making black-and-white cutoffs when it comes to certain mysterious phenomena, such as life and consciousness. People seem to want there to be an absolute threshold between the living and the nonliving, and between the thinking and the "merely mechanical,"…*

Thus a situation where free-will evolves in increasing effectiveness during the development of the brain satisfies the first of my criteria. Not only can the actions be free, but also the deliberation that resulted in those actions be free and the process to develop those deliberations be free etc. The fact that the chain of free-will disappears back into the internal evolutionary process can be expressed as a closure property.

The selective advantage that this feature confers upon us (as a species) is primarily that of external unpredictability (combined with an internal rationality). That is in a competitive environment, if an opponent can predict what you will do then that opponent would have a distinct advantage over you. Such competition in a social setting has been posited as one of the evolutionary selective factors that promoted intelligence in our species (Byrne and Whiten, 1988).

That unpredictability *can* be evolved has been shown by Jannink (1994). He developed a simulation with two separate populations which were co-evolved. The first of these populations was allocated fitness on the basis of the extent to which its programs successfully predicted the output of programs from the other and individuals from the second were allocated fitness to the extent that it avoided being predicted by individuals from the first population. Here the two populations are involved in a basic evolutionary 'arms-race'.

Thus the basic architecture I am suggesting is composed of the following elements:

- A framework for decision making processes;
- A population of processes within this framework;
- A way to construct new processes as a result of the action of existing decision making processes and the knowledge of the agent;
- A selection mechanism that acts to (1) select for those processes that tend to further the individual's goals and (2) to select against those processes that are predictable by others.

This evolutionary architecture is the basis for the suggested implementation. However, this architecture needs several more features in order to realise its potential. These are now discussed.

4 Open-ended strategy evolution

In a standard Genetic Algorithm (GA) following Holland (1975), the genome is a fixed length string composed of symbols taken from a finite alphabet. Such a genome can encode only a finite number of strategies. This finiteness imposes a ceiling upon the possible elaboration of strategy. This can be important where individuals are involved in the sort of modelling "arms-race" that can occur in situations of social competition, where the whole panapoly of social manouveurs is possible: alliances, bluff, double-crossing, lies, flattery etc. The presence of a complexity ceiling in such a situation (as would happen with a GA) can change the outcomes in a qualitatively significant way, for example by allowing the existence of a unique optimal strategy that can be discovered.

This sort of ceiling can be avoided using an open-ended genome structure as happens in Genetic Programming (GP) or messy genetic algorithms. Within these frameworks, strategies can be indefinitely elaborated so that is it possible that any particular strategy can be bettered with sufficient ingenuity. Here I use the GP paradigm, since it provides a sufficiently flexible framework for the purpose in hand. It is based upon a tree-structure which is expressive enough to encode almost any structure including neural-networks, Turing complete finite automata, and computer programs (Spector et al. 1999).

The GP paradigm means that the space of possible strategies is limited only by computational resources. It also has other properties which make it suitable for my purposes:

- The process is a path-dependent one since the development of new strategies depends upon the resource of present strategies, providing a continuity of development. This means that not only can completely different *styles* of strategy be developed but also different ways of approaching (expressing) strategies with similar outcomes.

- The population provides an implicit sensitivity to the *context* of action – different strategies will 'surface' at different times as their internal fitnesses change with the entities circumstances. They will probably remain in the population for a while even when they are not the fittest, so that they can 're-emerge' when they become appropriate again. Thus agents using a GP-based decision-making algorithm can appear to 'flip' rapidly between strategies as circumstances make this appropriate.

5 Meta-evolution

Such a set-up does mean that the strategy that is selected by an agent is very unpredictable; what the currently selected strategy is can depend upon the history of the whole population of strategies due to the result of crossover in shuffling sections of the strategies around and the contingency of the evaluation of strategies depending upon the past circumstances of the agent. However the *method* by which new strategies are produced is not dependent upon the past populations of strategies, so there is no backward recursion of the choice property whereby the presence of free choice at one stage can be 'amplified' in the next.

Thus the next stage is to include the operators of variation in the evolutionary process. In the Koza's original GP algorithm there are only two operators: propagation and tree-crossover. Instead of these two operators I suggest that the population of operators themselves are specified as trees following (Edmonds 1998). These operators are computationally interpreted so they *act upon* strategies in the base population to produce new variations. The operators are allocated fitness indirectly from the fitnesses of the strategies they produce using the "bucket-brigade" algorithm of Holland (1975) or similar (such as that in Baum 1998, which is better motivated).

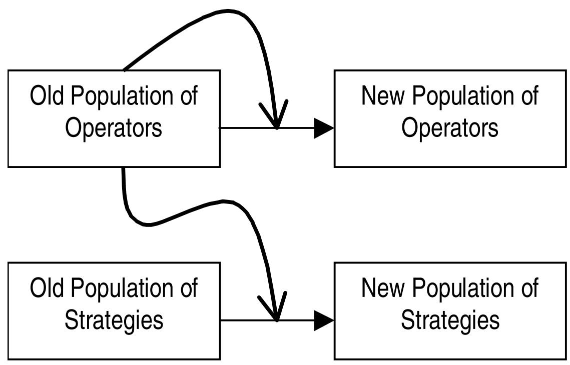

Figure 1. One step of a reflective meta-genetic process

To complete the architecture we set the population of operators to also operate on themselves in order to drive the production of new operators. Now the decision making processes (including the processes to produce the processes etc.) are generated internally, in response to the twin evolutionary pressures of deciding what to do to further the agents goals (in this case profit) and avoiding being predictable to other agents. This is illustrated in Figure 1.

6 Anticipatory rationality

If an agent is to reflectively *choose* its action rather than merely *react* to events, then this agent needs to be able to *anticipate* the result of its actions. This, in turn, requires some model of the world, i.e. some representation of the consequences of actions that has been learnt through past interaction with that world (either via evolution of the entity).

The models of the consequences of action are necessarily separate from the strategies (or plans) for action. It is possible to conflate these in simple cases of decision making but if an entity is to *choose between* plans of action with respect to the expected outcome then this is not possible. There is something about *rationality* which excludes the meta-strategy of altering one's model of the world to suit ones chosen strategy - the models are chosen according to their accuracy and relevance and the strategies are *then* chosen according to which would produce the best anticipated outcome according to the previously selected world model.

A reactive agent may merely work on the presumption that the strategies that have worked best in the past are the ones to use again. This excludes the possibility of anticipating change or of attempting to deliberately 'break-out' of current trends and patterns of behaviour.

Thus we have a process which models the consequences of action and one which models strategies for action. To decide upon an action the best relevant model of action consequence is chosen and the various strategies for action considered with respect to what their anticipated consequences would be if the consequence model is correct. The strategy that would seem to lead to the consequence that best fitted the goals would be chosen. This is illustrated in figure 2 below.

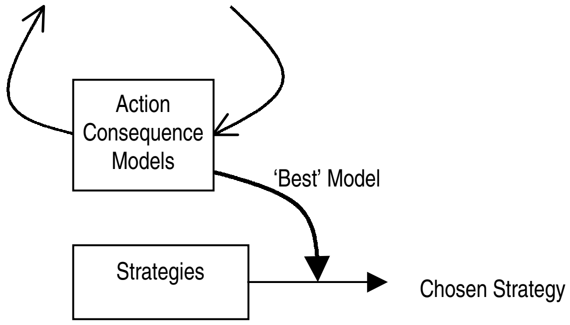

Figure 2. Using anticipation with strategy selection

7 Co-evolution

The next important step is to situate the above elaborated model of strategy development in a society of competitive peers. The development of free-will only makes sense in such a setting, for if there are not other active entities who might be predicting your action there would be no need for anything other than a reactive cognition. This observations fits in with the hypothesis that our cognitive faculties evolved in our species due to a selective pressure of social origin (Byrne and Whitten, 1988).

Thus we have a situation where many agents are each evolving their models of their world (including of each other) as well as their strategies. The language that these strategies are limited to must be sufficiently expressive so that it includes strategies such as: attempting to predict another's action and doing the opposite; evaluating the success of other agents and copying the actions of the one that did best; and detecting when another agent is copying one's own actions and using this fact to do what would help you. Thus the language has to have 'hooks' that refer to ones own actions as well as to other's past actions and their results.

In circumstances such as these it has been observed that agents can spontaneously differentiate themselves by *specialising* in different styles of strategies (Edmonds, 1999). It is also not the case that just because these agents are competing that they ignore each other. Such a co-evolution of strategy (when open-ended and resource limited) can result in the intensive use of the actions of others[3] as inputs to their own deliberation, but in a way that is unpredictable to the others (Edmonds, in press). So that the suggested structure for agent free-will can include a high level of social embedding.

[3] Including the communicative acts of others

8 Structuring the development of free-will within a society of peers

The final difficulty is to find how to structure this mental evolution so that in addition to maintaining the internal coherence of the deliberations and their effectiveness at pursuing goals and being unpredictable to others, the actions of the agent can be presented to others as rational and verified as such by those agents. This is in order to fulfil criterion (E) above.

This last criterion can be achieved if there is a normative process which specifies a *framework* of rationality which is not restrictive so that different deliberative processes for the same action can be simultaneously acceptable. The framework must be loose enough so that the *openness* of the strategy development process is maintained, allowing creativity in the development of strategies, etc. But on the other hand must be restrictive enough so that others can understand and empathise with the deliberative processes (or at least a credible reconstruction of the processes) that lead to action.

There are number of ways in which this framework could be implemented. I favour the possibility that it is the *language* of the strategies which is developed normatively in parallel with the development of an independent free-will. Thus the *bias* of the strategies can be co-evolved with the biases of others and the strategies developed within this bias.

9 Putting it all together

Collecting all these elements together we have the following parts:

- A framework for the expression of strategies which is (at least partially) normatively specified by the society of the entity.

- An internal open-ended evolutionary process for the development of strategies under the twin selective pressures of favouring those that further the goals of the entity and against those that result in actions predictable by its peers.

- That the operators of the evolutionary process are co-evolved along with the population of strategies so that indeterminism in the choice of the entity is amplified in succeeding choices.

- That models of the consequences of action be learned in parallel so that the consequences of candidate strategies can be evaluated for their anticipated effect with respect to the agent's goals.

Each of these elements have been implemented in separate systems, all that it requires is that these be put together. No doubt doing this will reveal further issues to be resolved and problems to be solved, however doing so will represent, I suggest, real progress towards the goal of implementing free-will.

10 Conclusion

Although it is probably not possible to implement the facility for free-will directly in an agent (i.e. by *designing* the detail of the decision making process)[4], I have argued that it is possible to implement a cognitive framework within which free-will can *evolve*. This seems to require certain machinery: an open-ended evolutionary process; selection against predictability; separate learning of the consequences of action; anticipation of the results of action and the evolution of the evolutionary process itself. Each of these have been implemented in different systems but not, as far as I know, together.

The free-will that results is a practical free-will. I contend that if the architecture described was implemented the resulting facility would have the essential properties of *our* free-will from the point of view of an external observer. Such a facility seems more real to me than many of the versions of free-will discussed in the philosophical literature, because it is driven more by practical concerns and observations of choice and is less driven by an unobtainable wish for *universal* coherency.

There are basically three possibilities: free-will is a sort of ‘magic’; it is an illusion; or it is implementable. I hope to have made the third a little more real.

Acknowledgements

Some of the cognitive models mentioned in this paper were implemented in the modelling language SDML[5]. SDML has been developed in VisualWorks 2.5.1, the Smalltalk-80 environment produced by ObjectShare. Free distribution of SDML for use in academic research is made possible by the sponsorship of ObjectShare (UK) Ltd. The research reported here was funded by the Faculty of Management and Business, Manchester Metropolitan University.

Philosophical appendix

There is no stopping some people philosophising, however inappropriate or unhelpful this is in particular contexts. Such people seem to think that it is both possible and useful to formulate generally and reliably true principles (i.e. principles completely without exceptions regardless of context) about the world using argument. For these people I *briefly* outline my position below, for full details they will have to come and argue with me.

- The "hard" deterministic thesis is untestable and has no practical consequences – the world is equally explained using it or otherwise since we can not rewind the world to see whether this thesis does in fact hold. The only consequences it can thus have is to sanction a normative claim about the use of the term "free-will" – this amounts to no more than a position that *given I conceive of the world as determined then there can not be anything denoted as an indeterministic process including free-will.*

- The above point can be demonstrated by considering the following thought-experiment: compare a human who had a ‘real indeterminism pump’ with an otherwise identical human with only a good ‘pseudo-random’ generator – there would be no testable or practical difference between them. The distinction is thus irrelevant except in how we conceive of our world.

- There is a lot of evidence against the hard deterministic thesis both at the micro-level (quantum effects) and at the macro level (that many complex systems are not determinable in practice).

- Any strengthening of the deterministic thesis to make it actually applicable (e.g. that given *almost* identical circumstances a certain identifiable system will exhibit the same behaviour) renders it false when applied to some systems – for example humans will not always exhibit the same behaviour in practically indistinguishable circumstances (even if they do not recall their previous decisions).

- I can see no reason why an *indeterministic* process has to be *arbitrary*[6].

- It is difficult to see how any conception of free-will that did not come down to the principles (A)-(E) above could have any realisable meaning.

- It is very difficult to see how the facility of free-will evolved in us as a species if it was not implementable and was inextricably linked with its practical consequences so it could be selected for.

---

[4] Indeed in another paper to be presented at the “Starting from Society symposium here at AISB2000, I argue that *intelligence* (as defined by the Turing Test) cannot be designed according to an explicit plan.

[5] Strictly Declarative Modelling Language - see http://www.cpm.mmu.ac.uk/sdml/

[6] although it possible that it is to all practical purposes *random* from the point of view of an external observer, depending upon the conception of randomness.

- It is much more useful (in the analysis of issues surrounding free-will) to consider the practical sources, advantages and consequences of different *kinds* of free-will as argued in Dennett (1984) and Sloman (1992).

Thus the practical, common-sense conception is a better representation of free-will than many of the idealised, philosophical characterisations of it. When involved in a process of implementation, it is wise to base one's work on the *best* representation available.

References

Baum, E. Manifesto for an Evolutionary Economics of Intelligence. In Bishop, C. M. (ed.), *Neural Networks and Machine Learning*, Springer-Verlag, 285-344, 1998.

Byrne, R. W. and Whiten, A. (eds.) Machiavellian Intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans, Oxford: Clarendon Press, 1988.

Dennett, D. C. Elbow Room: varieties of free-will worth having. Oxford: OUP, 1984.

Edmonds, B. Meta-Genetic Programming: Co-evolving the Operators of Variation. CPM Report 98-32, <http://ww.cpm.mmu.ac.uk/cpmrep32.html>, 1998.

Edmonds, B. Gossip, Sexual Recombination and the El Farol bar: modelling the emergence of heterogeneity. *Journal of Artificial Societies and Social Simulation*, **2**(3), <http://www.soc.surrey.ac.uk/JASSS/2/3/2.html>, 1999.

Edmonds, B. Capturing Social Embeddedness: a constructivist approach. *Artificial Behavior*, **7**(3/4), in press.

Harnad, S. Turing Indistinguishability and the Blind Watchmaker. In: Mulhauser, G. (ed.) *Evolving Consciousness*, Amsterdam: John Benjamins, in press.

Hofstadter, D. R. Analogies and Roles in Human and Machine Thinking, In *Metamagical Themas*, New York: Basic Books, 1985.

Holland J. H. Adaptation in Natural and Artificial Systems : an introductory analysis with applications to biology, control, and artificial intelligence. Ann Arbor : University of Michigan Press, 1975.

Jannink, J. Cracking and Co-evolving Randomizers. In Kinnear, K. E. (ed.) *Advances in Genetic Programming*, Cambridge, MA: MIT Press, 425–444, 1994.

Koza, J. R. Genetic Programming: on the programming of computers by means of natural selection, Cambridge, MA: MIT Press, 1992.

Palmer, R. G., Arthur, W. B., Holland J. H., et al. Artificial Economic Life: a simple model of a stockmarket, *Physica D* **75**:264-274, 1994

Sloman, A. How to Dispose of the Free-Will Issue. *AISB Quarterly*, **82**, Winter 1992-3, 31-32, 1992.

Spector, L., Langdon, W. B., O'Reilly, U-M., and Angeline, P. J. (eds.) *Advances in Genetic Programming, Volume 3*, Cambridge, MA:MIT Press, 1999.