

Implementing Free Will

A chapter proposal for “Visions of the Mind” in the form of an extended abstract, which is a complete rewrite of my AISB2000 paper

Bruce Edmonds
Centre for Policy Modelling
<http://cfpm.org/~bruce>

Modelling and Context

All usable modelling is context-dependent. When something interferes from outside the modelling context we often use a ‘proxy’ for this in the form of an effectively random input (e.g. a pseudo-random generator). This is often the best we can do since we cannot extend the model to capture what is beyond the modelling context, but a random source at least mimics the extra-contextuality of this interference. Thus *from within the context* causal factors are either (potentially) encodable as part of a model or, at best, thought of as random. However this does not *make* this interfering cause random in any other sense (i.e. in another or wider context).

A Functional Description of Free-Will

What common sense identifies as free will occurs in different degrees in different circumstances. It has evolved in us as a species (free-will is one thing that distinguishes us from bacteria). Thus it is likely to have given us selective advantage (at least, as individuals). I suggest that the properties of free will that are relevant are:

1. From the point of view of another the actions of the individual possessing free will are (at least somewhat) unpredictable.
2. One’s actions lead to one’s goals. That is, from an internal view the actions are consistent with achieving the goals.
3. When requested, the individual can produce an explanation for the (previously somewhat unpredictable) action which is acceptably rational. That is an account of ones decision process can be made in terms that are socially acceptable to other participants.

The advantage of these properties comes from the social context humans necessarily inhabit (‘necessarily’ because the survival of humans seems to come from their ability to inhabit many different ecological niches due to their social adaptivity). In a partially competitive social situation, where your competitors are trying to “guess” what you will do, there is obvious advantage in not being predictable by others. Yet at the same time one needs to perform actions that will further one’s own goals. Further, membership of many human social groups and institutions (in the widest sense) is often conditional on being able to demonstrate that one is rational (from the viewpoint of the others in that group) so that they can have some assurance that you will abide by the norms and rules of the group (e.g. incentives and sanctions will behave some sway on you). Thus simultaneously possessing abilities (1), (2) and (3) is advantageous for us humans (and to a lesser degree other highly social animals).

Criteria (1), (2) and (3) thus form our requirements for an implementation of free will. It is notable that these criteria are each about different contexts. (1) concerns only the external viewpoint of a competitor; (2) is only about the internal, cognitive context;

and (3) concerns the translation of the internal into the external context by an individual. Also not that all of these criteria are necessary:

A Proposed Mechanism for Free-will

The question is *how can these requirements be reconciled with a single mind*. We get a clue from ontogeny. What common sense identifies as free will appears in us during development (an adult has more free-will than a day-old foetus). This is not an instantaneous, all-or-nothing ability but one that develops in us over time. Once we have become an adult it is helpful to have created an internal context that is commensurable with other's internal contexts (or at least the social presentations of those internal contexts) but not be totally accessible to modelling from those other contexts. A process that is known to be able to *create* new contexts is that of evolution (in the widest sense). The hypothesis is that the brain has evolved so as to facilitate the evolution of free will over the development process. Thus the proposal is to implement free will using an evolutionary process, that is to adapt mechanisms from evolutionary computation to produce a model of the mind that meets criteria (1) and (2) (criterion (3) is more complex to achieve as it also requires social and communicative abilities, but mechanism to do this will be sketched).

The proposed mechanism is as follows: there is a 'space' of constructible strategies from a 'language' of steps, conditionals and actions; there is a current 'population' of strategies that are being evolved as the result of experimental variation of these and their evaluation (based upon the success of using the strategies or similar strategies); the language of these and many of the original archetypes for these strategies have a social origin, i.e. they are socially shared; the language must be suitably 'open-ended' (that is similar strategies in terms of effect must be expressible in many different ways) so that internal expressions of strategies will be different across individuals and so it is possible that developed variations result in different actions given the same circumstances; that the success of strategies will be according to the three criteria: unpredictability; effectiveness; and social accountability. This basic structure is augmented in two ways by adding the ability to anticipate the results of as strategies and thus allow the evaluation of strategies by whether they produced the expected results as well as the extent to which they furthered goals, and to also allow the (limited) co-evolution of the evolutionary operators themselves.

I claim that such a mechanism, when used in a socially embedded individual will satisfy the criteria above. It also answers the objection that free-will is only possible if the decision mechanism and the previous state is freely chose, for if you try and chase any particular decision backwards in time then you merely increase the difficulty of modelling it, so this becomes impractical. The roots of decisions are lost back in the evolutionary process – in a sense, this process can be seen as an way of amplifying the difficulty of modelling (from an external point of view) from small difficulties to insurmountable ones – or infinitesimal amounts of free-will up to effective amounts (the strangeness of the later analogy results from the attempt to impose a context-independent account on a process which creates a new context).

The Social and Cognitive Views

This picture of the relationship between the human mind and its social context as a (at least partial) explanation for human intelligence is the core of the "social intelligence" and "Machiavellian Intelligence" hypotheses. In particular, the latter version is

almost inevitable given the dual needs for making one's actions unpredictable before and yet revealed as furthering one's own ends after action. Thus the above implementation of free-will fits well with these hypotheses for a consequence of it is that it would occur and have meaning in a social context.

That the brain implements such a process (or something equivalent) is not clear. That it *could* implement such a process is indicated by the discovery that it does utilise evolutionary process as part of its functioning.

Philosophy

Such a proposal as this inevitably provokes many philosophical responses. They are mostly variants of an *a priori* conviction that free will is impossible, and thus my suggestion is inadequate. The most easily dismissed of these comes out of an assumption that the world (in some sense), and hence human decision making, is deterministic, despite the fact that evidence does not support this. A more sophisticated response is that the world is either deterministic or (as physics suggests) random, so human decisions are the same. However, as I argue at the beginning, this is a result of the context-dependency of our modelling, we use randomness for a model of what we can not model and we impute this upon the parts of human decision making we can not model. That we cannot model it is unsurprising since this is (part of) the purpose of free will – to separate the modelling from an internal and external point of view. This brings us to the most fundamental difficulty: it is part of the nature of philosophy to seek universal (non-context-dependent) models for the world leading to an, in practice, assumption that such a view is possible once the details of particular contexts are 'filtered out'. If the mechanisms of free will are based in their effectiveness at separating internal and external contexts (from the point of view of modelling) then its existence and the applicability of a philosophical approach are opposed (to the extent philosophy is formed of potentially non-context-dependent arguments or truths). One is then left with a choice: accept that such free-will can exist, for which there is some evidence (albeit most is anecdotal), or rely on the universality of philosophy (which is pure assumption since there can not be evidence for this).

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.
- Dennett, D. C. *Freedom Evolves*. London : Allen Lane, 2003
- Edelman, G. M. *Bright air, bright fire: on the matter of mind*. London : Penguin, 1992.
- Edmonds, B. *Meta-Genetic Programming: Co-evolving the Operators of Variation*. Invited paper in a special issue of *ELECTRIK on AI*, 9:13-29, 2001.

- 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), 323-348, 1999.
- Edmonds, B. Towards Implementing Free Will. , AISB'2000 symposium on "How to Design a Functioning Mind", Birmingham, April 2000 <http://cfpm.org/cpmrep57.html>
- Edmonds, B. (1999) The Pragmatic Roots of Context. CONTEXT'99, Trento, Italy, September 1999. *Lecture Notes in Artificial Intelligence*, 1688:119-132.
- Edmonds, B. (in press) The Social Embedding of Intelligence - Towards producing a machine that could pass the Turing Test. In Peters, G. and Epstein, R (eds.) *The Turing Test Sourcebook: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, Kluwer.
- 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 randomList. 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.
- 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.