The Social Embedding of Intelligence

Towards producing a machine that could pass the Turing Test

Bruce Edmonds
Centre for Policy Modelling,
Manchester Metropolitan University
http://bruce.edmonds.name

Key words: intelligence, social embedding, interaction, free will, empathy, self, emotion, Turing Test, learning, design, acculturation

Abstract: I claim that in order to pass the Turing Test over any period of extended time, it will necessary to embed the entity into society. This chapter discusses why this is, and how it might be brought about. I start by arguing that intelligence is better characterised by tests of social interaction, especially in open-ended and extended situations. I then argue that learning is an essential component of intelligence and hence that a universal intelligence is impossible. These two arguments support the relevance of Turing Test as a particular but appropriate test of interactive intelligence. I look to the human case to argue that individual intelligence utilises society to a considerable extent for its development. Taking a lead from the human case I outline how a socially embedded artificial intelligence might be brought about in terms of four aspects: free-will, emotion, empathy and self-modelling. In each case I try to specify what social 'hooks' might be required in order for the full ability to develop during a considerable period of *in situ* acculturation. The chapter ends by speculating what it might be like to live with the result.

Robert French (French 1990) rightly points out that the Turing Test is a test of human social intelligence rather than of a putative 'general intelligence' – I agree. The test is an inherently social one, which will necessitate the intelligence being embedded into the society in which it is being tested (in the case of the Turing Test, human society). This chapter will discuss why this is, how this might occur for an artificial intelligence and how this might enable the development of some of the human-like abilities and behaviour that would be necessary in order to pass the Turing Test. In particular I will look at: free-will; the sense of self; empathy and emotion. Before that I argue that there is no such thing as a general intelligence, and so the Turing Test is a very relevant test of intelligence as far as we are concerned. The reasons why a general intelligence is not possible will help motivate the following, more constructive discussion. I end by briefly speculating on how it might be like to live with the result.

1. SOCIAL TESTS FOR INTELLIGENCE

In this section I suggest that intelligence is usefully characterised by the ability to succeed with respect to one's goals in situations of extended interaction with other entities. This will lead back to the Turing Test as a special case. This is in contrast to the classic "constrained problem solving" paradigm, where one has to find the right sequence of discrete moves to obtain a known goal. Classic examples of constrained problems used in AI include the "blocks-world" puzzle, rectilinear mazes, the iterated prisoner's dilemma, and the travelling "salesman problem". Most of these puzzles have the following features in common: a sequence of discrete choices have to be made; there are a (usually small) finite number of pre-defined possibilities for the choices each turn; they are played off-line with respect to the world; and the goals are known to all. Such 'toy' problems test for only a very restricted aspect of intelligence and almost all researchers accept that such problems are an inadequate test of general intelligence. However such tests continue to dominate the literature – presumably there is an

assumption that these sorts of test are a sensible starting place, en route to a more substantial intelligence. However, one suspects that researchers often use such tests because they are easy to set up and they have not thought of any alternatives.

To illustrate some of the alternatives, consider the following situation: one has a series of candidates and one wishes to select the most intelligent of them with respect to some job (we will assume the candidates are human for the moment). How might one go about this task? Some possibilities are: interview them; look at what they have achieved in the past; read what people say about them; look at their qualifications; set them a challenge; and work with them over a trial period. Unless one wants an archetypal "egg-head" what you do *not* do is rely solely on their performance at solving a constrained puzzle. Even examinations, which come closest to the constrained problem-solving paradigm, usually require the ability to understand written text and write appropriate responses of one's own construction. For real job recruitment, people usually use a combination of evaluating past experience, taking references and conducting an interview, but there is general agreement that the only sure way to assess intelligence is to interactively work with the candidate over an extended period of time.

Unlike the constrained problem-solving paradigm, most situations concerning interaction between active entities allow for an indefinite number of possible actions. That is, the situation can not be meaningfully represented as a "game" with only a finite number of possible "moves". For example, if I was to try and catch a particular wild fox each night (and then release it ready for the next night) and I limited myself to a few strategies the fox would soon learn to avoid these. Rather, to be successful at this task I would probably be forced to continually innovate new approaches. This "open-ended" characteristic of the interaction is one of the features that makes this sort of situation a more appropriate test of intelligence than that of a "closed" game with a limited menu of choices. The necessity and type of such innovation indicates the intelligence of the fox: if a single, fixed strategy (that was potentially avoidable by the fox) worked, it would not indicate that it had much intelligence at all; if it was necessary to continually innovate in simple ways (e.g. move a trap to different positions but not change its nature) this would indicate at least some intelligence; if continual innovation in terms of the *type* of strategy was necessary then this would indicate a substantial intelligence (e.g. a low cunning); if whatever I did it evaded me then this might indicate that it was more intelligent than I was (in this context); and if it evaded me even when I employed outside help (consulting experts etc.) this might cause me to be so concerned as to suspect that there might be more behind this than was immediately apparent (e.g. unknown interference by another intelligent entity).

The above case is where the participants of the process have conflicting goals, resulting in a sort of cognitive "arms race". Conflict is not a necessary part of an interactive test of intelligence - it may be that the situation is one where the actors are co-operating. For example, in a game of "charades" the more inventive one is at finding ways of communicating information so that one's partner will understand, the better one is at the game. Charades has an obvious parallel with the situation where one person is attempting to communicate a new idea to another - the many cycles whereby questions are asked and assumptions checked and the flexibility in finding ways to "frame" the new idea with respect to ideas already known, require considerable intelligence.

The time-span of the interaction is another factor in such interactive tests for intelligence. While it is sometimes possible to succeed in such interactive situations using relatively shallow strategies over short periods of time, to succeed over in the longer term often requires increasingly sophisticated (and context-sensitive) strategies. It is relatively easy to convince someone at a party that one is something other that what one is because you would probably never meet them again. It is an altogether more difficult task to sustain such a deception if one meets them every day which would give them a chance to test out one's replies against the world. Similarly while relatively simple strategies might succeed in the Turing Test for a few sentences in specific and formulaic contexts (as in ELISA, (Weizenbaum 1976)), tying to maintain the illusion of humanity when the opponent can test the knowledge against others in society is an altogether tougher task (Dennett Daniel 1984). The Turing Test is certainly an open-ended test of interactive ability and could be conducted over a suitably extended period of time (what I called the "Long-term Turing Test" in (Edmonds 2000)). Passing the Turing Test necessitates the ability to successfully engage in a "cat and mouse" cognitive arms race similar to that with the fox

as well as the sort of communicative game evident in charades. Thus, despite its necessarily specific nature the Turing Test has a good claim to being one of the most relevant tests for the type of intelligence that most concerns us.

However the Turing Test is not only a suitable test of interactive ability, but also one that tests for an ability to imitate humans. This make it a particularly poignant test because the process whereby we might attribute intelligence to a program trying to "pass-off" as a human mirrors the process by which we attribute intelligence to each other. It seems obvious that this is partly why Turing devised such a test, for if anyone denied that a machine that passed the test was not intelligent, then this would undermine their attribution of each other's intelligence since that is made on the same basis. However, I can see no reason why the ability to imitate a human is necessary for intelligence – it is merely sufficient. However, I argue that many aspects of our intelligence are rooted in our interactive and social abilities. Thus to succeed at social interaction will require abilities that are closely associated with human abilities (even to the extent they are frequently used to *characterise* humanity). This is like an immigrant learning to cope in a new culture. They will need to acquire many skills and much knowledge specific to the culture in order to be able to build a successful life, but this does not mean they have to be able to "pass-off" as a native (a nearly impossible feat).

To summarise this subsection I have built up a picture of a test for intelligence rooted in interaction and which is: (1) over a substantial period of time which allows chances to test the content of interactions with the rest of the world (or society) and (2) which is open-ended, in that there are an indefinite number of possible interactions. The Turing Test is a specific case of this, but one which conflates the ability to succeed in complex interactive situations with the ability to "pass off" as a human.

2. THE IMPOSSBILITY OF A UNIVERSAL INTELLIGENCE

Two paradigms have heavily influenced the thinking on intelligence in AI over the last 30 years: the universal Turing machine (Turing) and the general problem solver of Ernst and Newell (Ernst and Newell 1969).

The first showed that there was (in theory) such a thing as a universal computational device which could be programmed to mimic any effective (symbolic) computation. In practice, this meant that one could mass produce the machines and leave their customisation to the programmer. While it does seem to be true that intelligence might be implementable using computation, this is very far from intelligence being reducible to computation. The second was a system of automatically decomposing higher goals into more tractable sub-goals so that one only had to specify a suitable top-level goal for the process to commence. This process of automatic sub-goaling is only possible when one can deduce the appropriate sub-goals from the goal and knowledge of the problem. This requires that one's knowledge is sufficient and consistent.

Thus the "universal" Turing machine is only useful when you know the correct program and the "universal" problem-solver can only operate when given consistent information that is sufficient to eventually decompose all its goals into solvable ones. These are, at best, narrow versions of "universality" – in everyday life one would almost never be able to use them without additional resources. Both systems require exactly the right amount of information – so that the solution (or result) is neither under- nor over-determined (this is, of course, enforced by the method of programming in the Universal Turing Machine case). If this is not the case these systems are simply inapplicable. Part of the problem lies in the lack of learning. In any but highly restricted and artificial problem-solving domains, learning is an integral part of intelligence. It enables a device to do something intelligent when it has either too little or too much information. Some examples of this are: when it is useful to perform an experiment upon some aspect of the environment to gain new information in order to make a decision; or when one detects a contradiction (or even incoherency) in one's beliefs and one has to decide which to ignore or change.

If one accepts that learning has to be a part of any autonomously applicable intelligence, then this has consequences for the intelligence's scope. A series of formal results in machine learning have

encapsulated what many in the field have suspected for some time – that there is no effective universal learning algorithm. Of these, one of the more recent is called the "No Free Lunch" theorems (Wolpert and Macready 1995). These show that over all possible search spaces no search algorithm does better on average than any other. This is a highly abstract result because most of the time we are not concerned with such spaces which are dominated by strange functions, such as those that are discontinuous everywhere. However it does show that to gain any effectiveness one has to exploit some knowledge of the search spaces that one is dealing with, even if this knowledge is minimal, for example that they are continuous or they have a finite number of maxima. An ability to learn more effectively than a random search is due to the exploitation of domain knowledge, which means that the algorithm is then limited (in terms of where it will be efficient) to the particular domain where that knowledge holds. In other words there is no universal and effective learning algorithm - the wider the domain of application the less efficient it might be and the narrower the scope the greater is the potential for efficiency.

This is a deep problem and not amenable to any easy fix. For example it may be thought that combining several different search algorithms so that the right one may be chosen by some control algorithm ((Minsky) suggests emotion plays this sort of role). This would then succeed in generalising the search algorithm and that this process of generalisation could be continued indefinitely. The problem with this is that ignores the cost of deciding which algorithm to use and the impossibility, in general, of choosing the right algorithm. All one has done is to switch from searching for solutions to searching for the right algorithm to search for solutions, which is not usually any easier than the original task.

If a universal learning algorithm is impossible and learning is an integral part of intelligence we must conclude that intelligence is also only effective and efficient when adapted to the circumstances it is used in. It is becoming increasingly clear that our intelligence is highly adapted to our circumstances (social, physical and biological). It does appear to be more generally applicable, in the short run, than those of other animals, especially when you consider inter-individual social processes of adaptation. However this marginal generality does not mean our abilities are completely general. The apparent generality comes at a cost – it is by no means clear that this is also the case in the long run or during abnormal events. Our socially oriented intelligence does seem to allow us to occupy a variety of ecological niches on earth by cultural rather than biological adaptation, but the same cultural abilities are also our weakness and could easily end up destroying us as a species.

One suspects that the underlying reason why it has been assumed that a universal intelligence is possible is anthropocentrism: humans like to assume that they are in possession of such an intelligence. The assumption seems to be that in all important aspects that our cognitive abilities are universally applicable and that any limitations (typically put down to "mere" memory and processing capacity limitations) can be overcome, e.g. by the use of tools such as: writing or computers. This strikes me as pure hubris: not long ago we assumed we were physically in the centre of the universe, we now seem to be making a similar sort of assumption in terms of our cognitive abilities.

Robert French (French) drew an analogy between trying to imitate the flight of a seagull (for an observer watching on radar) and trying to pass the Turing Test. He suggests that just as the flight of a seagull is only a particular type of flight, the Turing Test tests for only a particular type of intelligence - human social intelligence. However, this merely builds in the assumption that intelligence *can be* meaningfully generalised by choosing a situation where we happen to know there is an appropriate generalisation for us - flight. The analogy has apparent force by conflating abstraction and generalisation - flight can be both generalised and abstracted from the seagull's example and intelligence can be abstracted from the Turing Test, but this does not follow that it is possible to meaningfully to generalise from that which is tested by the Turing Test to a "universal intelligence".

3. THE SOCIAL CONSTRUCTION OF INDIVIDUAL HUMAN INTELLIGENCE

If one accepts that the ability to choose or construct actions that further one's goals in complex, extended, open-ended and social situations, is an appropriate characterisation of intelligence then this will have consequences for how such an intelligence can be brought about. The human case provides some clues as to how this might happen. It seems that human intelligence is heavily dependent upon society for its development. In this section I look at some of the evidence and arguments for this including: language; the ability to participate in the social web of co-operation and competition; the presence of exploitable information resources in society and the phenomena of autism.

A major component of human intelligence is language. Without language humans are not able to effectively co-ordinate action in order to achieve goals, without language humans could not discuss problems to brainstorm, criticise and compare solutions, and without language we would we not be able to reason so effectively. For example, it is almost inconceivable that a human without mastery of a sophisticated language could perform abstract mathematics. Full languages are not pre-programmed into our brains by our genes. To acquire a full language it is necessary for an individual to be socially immersed in a linguistic environment. The ability to make use of such an environment in order acquire language is genetic, but the language itself is largely learnt *in situ*. The way our linguistic intelligence is developed suggests a model for the development of social intelligence in general. That is, the individual is pre-programmed with certain abilities, biases and innate knowledge. These 'hooks' then allow the rapid learning of the contingent knowledge and skills through interaction in the society. This approach allows for the contingent information to be adapted to the circumstances of the society.

The recently proposed "social intelligence" (Kummer, Daston et al. 1997) and "Machiavellian intelligence" (Byrne Richard and Whiten 1988; Byrne Richard and Whiten 1997) theses put forward the theory that substantial aspects of our intelligence evolved because its possession conferred social advantage. The idea is that our extensive intelligence is primarily evolved in order to keep our place in the social order and to manage the intricate web of co-operation and competition that this involves. The idea is that the evolutionary advantage our intelligence comes from our ability to develop different cultures suited to different niches and to socially organise to exploit these.

If this is the case (and it would be odd if none of our intelligent capacity has been shaped by evolutionary pressures that are socially grounded), and given the intricacy of our present society (which presupposes the possession of individual intelligence) then it seems likely that our intelligence and our society have co-evolved. If this is the case then one would expect that substantial aspects of our intelligence have evolved to 'fit in' with our society (and vice versa). It is certainly difficult to argue from single cases, but the fact that the only species to evolve a sophisticated intelligence has also evolved a sophisticated society can not be totally ignored.

One aspect of a society of roughly commensurate social entities that is almost inevitable is that it will quickly develop so as to be more complex than any single entity can completely model. This is especially true of a society where there is sometimes advantage in 'out-guessing' the actions of the other entities, in which case a sort of cognitive 'arms-race' can develop which, in its turn, makes the prediction and comprehension of the society even more difficult.

Given that our society will be more complex than we can ever understand, there will probably be societal processes that perform computations that we can not do individually. It would be strange if some of these informational and computational resources (i.e. the 'results' of the societal computation) were not accessible to some of the constituent entities some of the time. Given this availability it would also be odd if these resources were not exploitable by the entities it is composed of. Hence one would expect entities that were situated in such a society to evolve ways of exploiting such resources (indeed some simulations do suggest this (Edmonds 1999)). If this were the case, then we would expect that we would posses some faculties, usually attributed to our 'intelligence', that were evolved to use such resources and save ourselves (individually) considerable time and effort. This utilisation of information resources in society is analogous to the sampling of the immediate physical environment for information about our position etc. rather than relying for the detail from an internal 'map'

(Brooks). Thus one would expect that our brain has not only evolved because it allows the creation of culture but also that it would have evolved to exploit this culture.

One piece of evidence about the importance of society to individual intelligence is the phenomenon of Autism. Since the 1940s autism has been known as a syndrome which involves, among other features, the striking lack of social competence. A variety of explanations have been discussed, among them the widely discussed 'theory of mind' model which conceives of autism as a cognitive disorder (Baron-Cohen, Leslie et al. 1985), and, a more recent explanation given by (Hendriks-Jansen 1997). He hypothesises as the primary cause early developmental disorders which prevent the child and its caretakers to 'get the interaction dynamics right', the interaction dynamics which normally scaffold the development of appropriate social interactions in the sense of situated dialogues between infant and caretaker.

The importance of interaction dynamics are also part of the explanation given in (Dautenhahn 1997) which suggests a lack of empathic processes which prevent the child developing 'normal' kinds of social action and interaction. People with autism never develop into social beings as we expect of 'normal' people. Although some of them show high intelligence in non-social domains, they are never able to communicate and interact successfully with other people. They are not able to understand the social world around them, which therefore often appears as scary and unpredictable. This deficit influences their lives to the extend that they often are not able to lead an independent life, in this way clearly demonstrating the central role of sociality in practical intelligence. This gives evidence that socially situated aspects of intelligence do not merely provide an important add-on to other faculties of intelligence (like spatial thinking or mathematical reasoning), but that human intelligence (its development and expression) is embedded (and embodied) in a social being, and can in this way not be separated from non-social kinds of intelligence. The existence of autism shows two things: that there are different types of intelligence (e.g. mathematical and social); and that social intelligence is critical to survival in humans.

It often seems to be assumed that social intelligence is merely the application of general intelligence to the case of social interaction. However, the arguments and evidence above tend to indicate that there is another possibility – that a large part of our intelligence is to facilitate and exploit the social processes that give our species evolutionary advantage, and that our ability to use our intelligence for other ends is partly an offshoot of our social intelligence. In other words one of the main aspects in which our intelligence has specialised is that of managing our social interaction.

The development (and to a lesser extent the application) of our individual intelligence is thus dependent upon its development in a social context. This requires a considerable period of *in situ* training and development and is in sharp contrast to the way in which we tend to develop artificial intelligences. In AI the analytic design stance predominates – the required ability is analysed and then explicitly designed into a set of programs and hardware. If we value intelligence that is commensurate with human intelligence in terms of its social ability, then there is a real possibility that such a stance might be inadequate to the task.

4. TOWARDS PRODUCING A SOCIALLY EMBEDDED INTELLIGENCE

Passing the Turing Test (and especially the long-term Turing Test), requires that the entity is effectively embedded in human society, because this is the only way in which an entity can have access to the wealth of contingent and context-dependent information that is necessary to pass-off as human. In the analogous situation of the immigrant trying to assimilate into a new country, preparation before-hand by reading books about the country would not be sufficient to pass off as a native – rather it takes at least decades of immersion before this is possible. It is easier to learn to merely interact successfully in that culture, a process that would probably take years rather than decades, but this requires that one has already learnt to act successfully in *a* human culture. It takes a child decades before it is fully competent to interact in its first culture, and it is only because the

knowledge and skills gained is transfereble to a different human culture that a later, relatively rapid, viable application to a new culture is possible.

In this subsection I consider four aspects of human intelligence that are probably necessary for interactive intelligence. Each requires the social embedding of intelligence, albeit in slightly different ways. The aspects are: free-will; appropriate emotional responses; empathy; and self-modelling. These are not sufficient to pass the Turing Test (or similar interactive test) but are probably necessary.

In each case the strategy is the same: provide the cognitive processes and 'hooks' that would allow the development of the aspect during considerable *in situ* acculturation. I speculate about how the ability might arise and hence what 'hooks' and social environment might be necessary for their development. Although in each case some aspects of these mechanisms have been explored by me and others, these suggestions are necessarily highly speculative. However, it does appear to me that these abilities are deeply rooted in social processes for their fruition, so I would expect that some comparable set-up and process would do the job if these suggestions prove inadequate.

It must be said that very little is known about what kinds of abilities are necessary for and result in what kind of interaction. The distributed artificial intelligence and social simulation communities have made a start, but there is still much to do. Thus, although I do try to base the suggestions I make upon what I observe or theorise is necessary for these abilities to develop, they have only been tested in rudimentary and fragmentary ways. Also there are many other important aspects of human intelligence that are, no doubt, necessary, that I have not considered here, including: natural language, context-awareness and analogical thinking.

4.1 Free Will

I am not going to discuss the philosophical aspects of free-will here, but rather the aspects that are practically relevant to social interaction, and how these aspects might be brought about. From this perspective an entity with free-will has the following key properties:

- *firstly*, that some of its actions are not predictable before hand – a part of this is that given similar circumstances and history the entity may behave in different ways;
- but *secondly*, that the action is rational, by which I mean that it furthers the entity's goals – sometimes it is also necessary that this rationality is socially apparent (or can be socially demonstrated).

These are largely a question of effective power. In competitive situations it is often to an entity's advantage to be somewhat unpredictable, so that one's competitors will not be able to completely predict what you will do, even if they have frequently observed you in similar circumstances. On the other hand it is advantageous that ones actions further one's goals. Clearly these are not easy requirements to simultaneously satisfy.

Another aspect of personal power is that it is often dependent upon being a member of social groups, institutions and processes and this membership is often dependent upon being able to demonstrate that one is rational (with respect to acceptable goals). This reflects the fact that these institutions etc. depend upon constraints and inducements being effective on its members' actions – if a member is not rational then there will be no way for the institution to promote or enforce these constraints upon that member and the inducements may be ineffective. If one fails to convince others of one's rationality one can lose social advantage, either due to one's action being anticipated and countered or due to one being ineligible to participate in the social institutions that give access to socially produced goods (in the widest sense).

In the Turing Test one of the ways in which one might judge whether the entity was human or not is whether its responses to the same conversational sequence results in an unpredictable reply that is, afterwards, understandable in terms of imputable goals.

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

that the result is a will which is internally coherent in furthering its goals but yet not effectively predictable from its circumstances.

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

Perhaps the problem is the seeming need that people have of making black-and-white cut-offs 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 development get around the requirement for prior free-will. 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 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 fits in well with the social intelligence hypotheses mentioned above (Byrne Richard and Whiten 1988; Byrne Richard and Whiten 1997) since unpredictability can give social advantage and hence be selected for by evolution. That such an effective unpredictability *can* be evolved has been shown by Jannink (Jannink 1994). He developed an algorithm where two separate populations 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 second 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'.

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

- An expressive framework for expressing strategy forming processes;
- A population of such 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 entity;
- 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.

4.1.1 Open-ended strategy evolution

In a standard Genetic Algorithm following (Holland 1992), 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 panoply of social manoeuvres is possible: alliances, bluff, double-crossing, lies, flattery etc. The presence of a complexity ceiling in such a situation (as

would happen with a genetic algorithm) 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 (Koza 1992; Koza 1994) or messy genetic algorithms (Goldberg, Deb et al. 1989). 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 genetic programming 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 (Koza 1992; Koza 1994). Using the genetic programming paradigm means that only the available computational resources limit the space of possible strategies. It also has other properties which make it suitable for this purpose:

- 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 entities using a evolutionary decision-making algorithm can appear to 'flip' rapidly between strategies as circumstances make this appropriate.

4.1.2 Meta-evolution

Such a set-up does mean that the strategy that is selected by an entity 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 entity. 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 my next suggestion is to include the operators of variation in the evolutionary process. In the Koza's original genetic programming algorithm there are only two operators: propagation and tree-crossover. Instead of using only these operators I suggest that there be a whole population of operators that are themselves specified as trees following (Edmonds 2001). These operators can be computationally interpreted so they *act* upon strategies in the base population to produce new variations (instead of crossover). The operators are allocated fitness indirectly from the fitnesses of the strategies they produce using the "bucket-brigade" algorithm of Holland (Holland 1992) or similar (such as that of Baum (Baum), which is better motivated).

To complete the architecture we arrange it so that the population of operators also operates on their own population 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 entities goals (in this case profit) and avoiding being predictable to other entities. This is illustrated in figure 1.

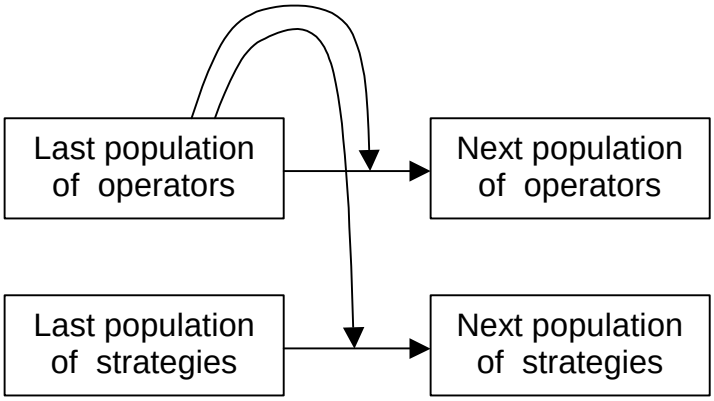

Figure 1. One step of a meta-genetic evolutionary process

4.1.3 Anticipatory rationality

If an entity is to reflectively choose its action rather than merely react to events, then this entity needs to be able to anticipate the result of its actions. This, in turn, requires some models 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 or a cognitive process). This has been called 'anticipation', and was first noticed by (Tolman), suggested as part of the schema in (Drescher), and included in evolutionary systems in (Stolzmann 1998).

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 seems to limit the meta-strategy of altering one's model of the world to suit ones chosen strategy – the models are chosen according to their accuracy in anticipating the effect of action (as well as their relevance) and the strategies are then chosen according to which would produce the best anticipated outcome according to the previously selected model. This is in contrast to a purely reactive entity that may work on the presumption that the strategies that have worked best in the past are the ones to use again. A reactive strategy excludes the possibility of anticipating change or being able to deliberately 'break-out' of current trends and patterns of behaviour. Thus what is proposed is a process that 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 action consequence model is correct. The strategy that would seem to lead to the consequence that best fitted the goals is chosen. This is illustrated in figure 2 below.

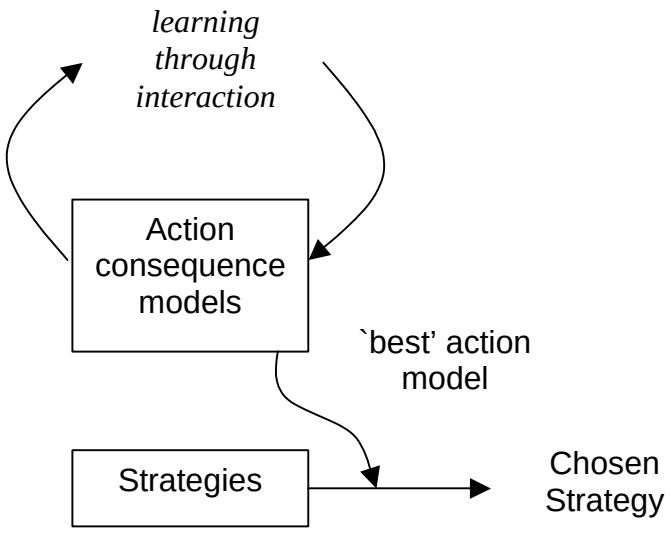

Figure 2. Using anticipation with strategy selection

4.1.4 Co-evolution

The next important step is to situate the above set-up 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.

Thus we have a situation where many entities are each evolving their models of their world (including of each other) as well as their strategies. The language that these strategies are expressed in needs to 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 entities and copying the actions of the one that did best; and detecting when another entity is copying one's own actions and using this fact. 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 entities can spontaneously differentiate themselves by specialising in different styles of strategies (Edmonds 1999). It is also not the case that just because these entities 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 as inputs to their own deliberation, but in a way that is unpredictable to the others (Edmonds 1999). So that the suggested structure for entity free will can include a high level of social embedding.

4.1.5 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 entity can be presented to others as rational and verified as such after the case by those entities.

This can be achieved if there is a normative process that specifies a framework of rationality that is not overly 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. On the other hand, it 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 an 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.

4.1.6 Putting it all together

Collecting all these elements together we have the following parts:

1. A framework for the expression of strategies which is (at least partially) normatively specified by the society of the entity.
2. 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.
3. 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.
4. 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 entity's goals.

Each of these elements have already 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 bringing about a practical free-will.

4.2 Emotion

Human emotion is a complex phenomenon (Sloman): it is partly innate and partly learned; it seems to have a pervasive affect on almost all aspects of cognition, including perception and action; it is associated with physiological affects, interpersonal signalling and phenomenal experience; it is affected by our thoughts, our environment and the emotions of others; some of our emotions seem to be shared with animals and others not.

Even if, as some have suggested, emotion is grounded in our physiology and the extent to which we achieve our goals (Rolls 2000), it then acquires layers of extra function through its social use. For example, it may be that anger has a basic role in triggering physiological changes to prepare us for action (e.g. fight) when we are threatened. However, these physiological changes are detectable by others, who may then decide to avoid a fight by backing down. Given this backing down, becoming angry might be an effective way of getting one's way without fighting, so one might learn to trigger anger in oneself in the absence of any threat. In another iteration of this cycle: if others observe one is using anger in this way they might learn to automatically meet anger with anger in situations where a fight is unlikely to make it clear that the person will not simply get his own way by being angry. Thus anger can be more important as a social signalling device than as a trigger for physiological change. It seems likely that it is by this sort of process that different cultures have developed different patterns of emotional response and display. Once brought up in one culture it is not possible to easily change one's emotional habits and expectations.

Here the idea is that, in practice, emotions are a special type of action – an action whose effect is on one's own body and brain. An essential part of the emotion-action is that it affects the way one considers and evaluates one's own models and strategies. Thus fear might release chemicals to facilitate our body for detection of danger and flight but it will also cause us to evaluate our ideas and perceptions pessimistically. These emotion-actions will be triggered in a similar way to other actions – sometimes it will be a simple and unconscious reaction (as might happen when we are woken by a loud and unexpected noise), and sometimes as a result of conscious deliberation (e.g. when we imagine what a teacher might say to us when it is revealed we have not done our homework).

Thus, in this proposal, the emotion acts upon the mechanism that evaluates models or strategies, and this, in turn, effects the model or strategy that is selected as the best (i.e. most appropriate under the circumstances). Thus the emotion does not determine the action but biases the selection of an action. Emotion-actions can be seen as a 'short-cut' concerning the feedback to action selection that has been learned over the course of evolution so that we do not have to learn it individually. If the

strategy learning process was implemented with an evolutionary algorithm, the emotion would change the evaluation of candidate strategies in the population of possible strategies being developed. This set-up is illustrated in figure 3.

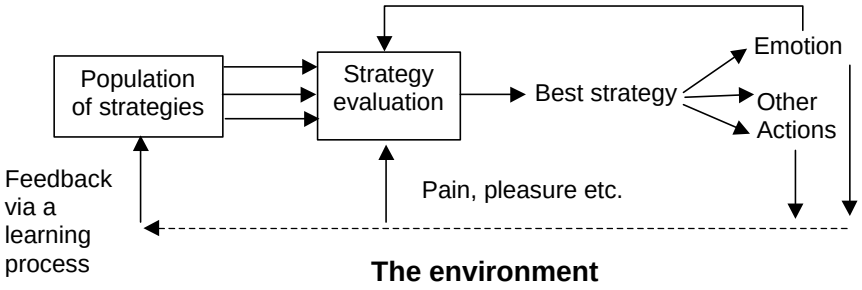

Figure 3. Emotion acting to affect the evaluation of models and strategies

Clearly the emotion-actions we have evolved relate to different situations and needs where the emotion will give adaptive advantage: fear to a threat, protectiveness to child-rearing etc. Many of these are as appropriate for an artificial entity as for humans. However these emotion-actions are elaborated and moderated in different ways for different circumstances. Some emotions will be moderated to suit the environment that the society inhabits. For example anger may be suppressed or redirected in societies where a high degree of co-ordination is necessary and where there is no significant physical threat from others. Other emotions may be moderated by fashion, e.g. crying in men.

Putting emotions into the learning loop of the entity, so that it biases the learning and evaluation of possible actions, allows it to be elaborated and moderated as a result of learning by the individual. Then it becomes possible to learn to choose an emotion because it is useful to do so, just like any other action. One can also (with more difficulty) learn to suppress an emotion, for example one can decide not to be angry (or develop one's anger). Of course, just as with other actions we can be surprised and take an action before we have thought about it, but this does not mean that such actions are always instinctual or reactive.

Sharing the same emotional machinery means that we know quite a lot about the likely emotions in others. This knowledge is only reliable for prediction in extreme cases, or when it will be difficult for the person to reflect (e.g. when woken in the night or drunk). At other times we know that the emotions can be considerably elaborated or moderated by culture and the individual, in which case it is no more predictable than other actions. However when the emotion does occur we can understand and empathise with it – this is the subject of the next subsection.

4.3 Empathy and Self-modelling

Human empathy involves being able to share the emotions of others. This seems to include (at least) the following steps:

1. *understanding* the emotions of another from their actions and situation;
2. *imagining* oneself experiencing the emotions of another based upon our understanding;
3. *experiencing* the emotions of another based on our imagination of them.

Other aspects of empathy might include: invoking one's own emotions in others and detecting whether other's emotions are real or feigned. Thus although empathy can be a powerful means of expressing sympathy it goes beyond sympathy, for sympathy does not (necessarily) include the duplication of other's emotions.

Each of the three steps listed above necessitates different cognitive abilities. Understanding another's emotions (step 1) necessitates that one has an adequate model of other people's emotions; imagining their emotions (step 2) that one is able to map one's model of other's emotions onto one's own; and experiencing them (step 3) that this mapping can trigger one's similar emotions in oneself. Of course empathy can work both ways – once one starts experiencing the emotions of another (e.g.

fear) this might reinforce the emotion in the person it originated in (or it may invoke it in a third person).

The ability to trigger similar emotions to someone else's in oneself implies a high degree of commonality between people. Since some of this commonality is culture-specific it must be learnt during development. Here I suggest a process whereby at an early stage an entity co-develops its self-model alongside its model of others. Later a differentiated self-model might be developed from this common base to suit the needs and experience of the individual.

I outline a model of how the self-model might be developed. This attempts to reconcile the following observations and theories about the human sense of self:

1. That the self is only experienced indirectly (Gopnik 1993).
2. That a self requires a strong form of self-reference (Perlis 1997).
3. That aspects of the self are socially constructed (Burns and Engdahl 1998).
4. "Recursive processing results from monitoring one's own speech" (Bridgeman 1992).
5. That one has a "narrative centre" (Dennett 1989).
6. That there is a "Language of Thought" (Aydede 1999) to the extent that high-level operations on the syntax of linguistic production, in effect, cause other actions.

The purpose of this is to approach how we might provide the facilities for an entity to construct its self-model using social reflection via language use. If the entity's self-model is socially reflective this allows for a deep underlying commonality to exist without this needing to be prescribed beforehand. In this way the nature of the self-model can be developed in a flexible manner and yet there be this structural commonality allowing empathy between its members.

This model is as follows:

1. There is a basic decision making process that acts upon the perceptions, actions and memories of the entity and returns decisions about new actions (this would include changing the focus of one's perception and retrieving memories).
2. The entity attempts to model its environment by a learning process that does not have direct access to the workings of this basic process but only of its perceptions and actions, past and present.
3. The modelling process seeks to model its environment, including the other entities it can interact with. It also attempts to model the consequences of its actions (including speech acts).

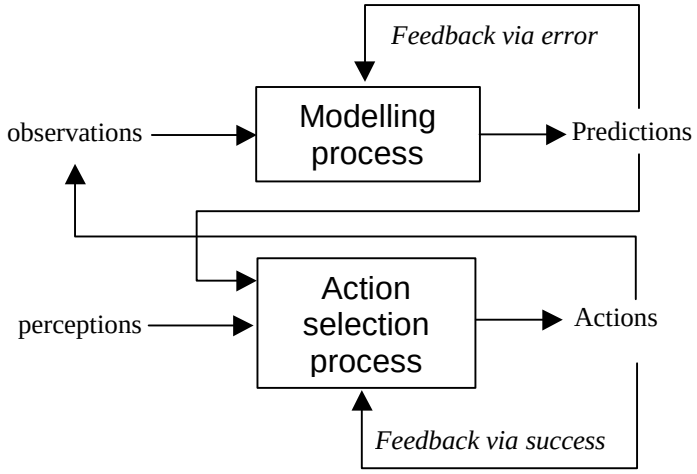

Figure 4. Separate modelling and decision-making processes feeding back into each other

4. The modelling process picks up and tries out selections of the communications it receives from other entities and uses these as a basis (along with observed actions) for modelling the decisions of these entities.
5. The action selection mechanism becomes adapt at using communication acts to fulfil its own needs via others actions using its model of their decision making processes (utilising the *results* of the modelling process in terms of predictions of their actions).

6. Using the utterances it produces it learns to model itself (i.e. to predict the decisions it will make) by applying its models of other entities to itself by comparing its own and others' acts (including communicative acts). The richness of the language allows a relatively fine-grained transference of other's decision-making processes onto itself.
7. It refines its model of other entities' actions using its self-model and its self-model from its observation of other's actions. Thus its model of other's and its own cognition co-evolve.
8. Since the model of its own decisions is made through language, it uses language to implement a sort of high-level decision-making process – this appears as a language of thought.

The key points are that the basic decision making process are not directly experienced; but rather the entity models others decision making using their utterances as fine-grained indications of their mental states (including intentions etc.); and then models its own action selection mechanism by applying its model of others to itself (and vice versa). This seems to be broadly compatible with observations in (Aydede and Güzeldere forthcoming).

4.3.1 General consequences of this set-up

The important consequences of this set-up are:

- The fact that models of other entities and self-models are co-developed means that basic assumptions about self and other's cognition are similar.
- The fact that an expressive language has allowed the modelling of others and then of its self means that there is a deep association of self-models with this language.
- Communication has several sorts of use: as a direct action intended to accomplish some goal; as an indication of another's mental state/process; as an indication of one's own mental state/process; as an action designed to change another's mental state/process; as an action designed to change one's own mental state/process; etc.
- Although the modelling processes do not have access to the basic decision making processes they do have access to and can report on their self-model which is a model of their decision-making. Thus, they do have a reportable language of thought, but one which is only a good approximation to the underlying basic decision making process.
- The model allows social and self reflective thinking without regression limited only by computational resources and ingenuity – there is not problem with unlimited regression, since there is no direct access between the modelling process and the action selection process.

4.3.2 Towards producing self-modelling entities

Working from the above model gives enough information to suggest the outline of a set-up which might bring about self-modelling. The basic requirements for this are:

1. a suitable social environment (including humans);
2. sufficiently rich communicative ability – i.e. a communicative language that allows the fine-grained modelling of others' states leading to action in that language;
3. a general anticipatory modelling capability;
4. an ability to distinguish the experience of different types, including the observation of the actions of others; ones own actions; and other sensations;
5. need to predict other's decisions;
6. need to predict one's own decisions;
7. ability to reuse model structures learnt for one purpose for another;

Some of these are requirements upon the internal architecture of an entity, and some upon the society it develops in. I will briefly outline a possibility for each.

The entity will need to develop two sets of models.

1. A set of models that anticipate the results of action, including communicative actions (this roughly corresponds to a model of the world). Each model would be composed of several parts:
    - a condition for the action

- the nature of the action
- the anticipated effect of the action
- (possibly) its past endorsements as to its past reliability

2. a set of models of strategies for obtaining its goals (this roughly corresponding to plans); each strategy would also be composed of several parts:
   - the goal
   - the sequence of actions, including branches dependent upon outcomes, loops etc.
   - (possibly) its past endorsements as to its past success

These can be developed using a combination of anticipatory learning theory (Hoffman 1993) as reported in (Stolzmann 1998) and evolutionary computation techniques. Thus rather than a process of inferring sub-goals, plans etc. they would be constructively learnt (similar to that in (Drescher 1991)). The language of these models needs to be expressive, so that an open-ended model structure such as in genetic programming (Koza 1992) is appropriate, with primitives to cover all appropriate actions and observations. However direct self-reference in the language is not built-in. The language of communication needs to be a combinatorial one, one that can be combinatorial generated by the internal language and also deconstructed by the same.

The social situation of the entity needs to have a combination of complex co-operative and competitive pressures in it. The co-operation is necessary if communication is at all to be developed and the competitive element is necessary in order for it to be necessary to be able to predict other's actions (Kummer, Daston et al. 1997). The complexity of the co-operative and competitive encourages the prediction of one's own decisions. A suitable environment is where, in order to gain substantial reward, co-operation is necessary, but that inter-group competition occurs as well as competition for the dividing up of the rewards that are gained by a co-operative group.

Some of the elements of this model have already been implemented in pilot systems (e.g. (Drescher 1991); (Edmonds 1999); (Stolzmann 1998)).

4.3.3 Back to Empathy

If one has a situation where entities' self-models have substantial commonality with each other and emotions are to an extent elaborated and moderated by culture-specific learning, then empathy is possible. The entity could *understand* other's emotions due to the commonality of the basic emotion-action's purpose and affect as well as the fact that it has extensively modelled the emotions of others. It could *imagine* these emotions because its model of its own emotions was co-developed with its model of other's emotions. Finally, it could *experience* these emotions having learned to trigger the corresponding emotions in itself in the appropriate social circumstances.

5. LIVING WITH THE RESULT

In considering what it will be like to live with the result I will consider two scenarios: the first is the situation where intelligences that can successfully interact in complex, social situations are developed; the second is the case where such intelligences were able pass the Turing Test (or similar). The basis of this speculation is an analogy between different intelligences interacting in an information "ecology" and species interacting in a biological ecology.

In the first case, there would be new intelligences that had the capability of interacting fully with each other and with humans, but are not necessarily similar enough to humans so that they could pass the Turing Test. At the moment the information ecosystem is dominated by a single type of intelligence: human intelligence. In this future, there might be many "species" of intelligences dealing with information. In a biological ecosystem competition for resources can result in the specialisation of species so as to avoid direct competition. This is possible as long as there is a diversity of niches in the ecosystem and the species are sufficiently different. In this case each species will have comparative advantage over other species in different circumstances, so that no one species dominates everywhere. If there is no such thing as a 'general species' then species are wiped out only if a

resource that they rely upon is used up or destroyed by another. If, as I have argued, a general intelligence is impossible and there are many, essentially different, ways in which one can exploit information then the creation of new intelligences might not lead to our eclipse, but it might well mean that we have to specialise in those areas where our kind of intelligence gives us comparative advantage. This would mean that we were not competitive in certain arenas, in the same way that most of us could not compete with professional musicians or sumo wrestlers. The difference would be that we could not pretend to ourselves that we had the potential to compete – the limitations of our intelligence would be thrown into sharp relief, by entities that are not human. When another human succeeds in a way we can not, we can gain comfort by imagining ourselves in their shoes or mentally associating ourselves with them and hence 'sharing' their success. With an artificial intelligent entity, this may be more difficult and so their relative success might be more difficult to bear. Historically people have resented most the newcomers that are least similar to themselves.

Sometimes the emergence of a new species creates new niches for other species (e.g. mollusc shells used by hermit crabs) - or even opportunities for symbiosis. In a similar way it may be that new intelligences create new opportunities for human intelligence in ways that we can not foresee.

If new intelligences are brought about that are sufficiently close to human intelligence to pass the Turing Test, they may be much more difficult to live with. The closer the intelligences are to human intelligence, the more directly they will compete with humans and the less likely it is that the new and human intelligences will be able to specialise to exploit different types of interactive opportunity or informational resource. This may mean that the informational niches that humans are left with are relatively restricted and unimportant. In the UK grey squirrels and red squirrels inhabit the same niche, so that it is inevitable that (in the absence of outside intervention) that one species will be wiped out (as has occurred except in a few local areas where the grey squirrels are culled by humans to protect the red).

On the other hand, if they are similar to us (as in some way they must be to pass the Turing Test), maybe our demise will be softened by their similarity with us – we can pass away with the comforting thought that our image lives on in them (until they reinvent it in their own society).

6. CONCLUSION

The Turing Test happens to be quite a good test for interactive intelligence of a particular type - far better than any constrained puzzle or 'toy problem'. It does not (necessarily) demand the physical embodiment of the entity but it does necessitate its social embedding. Several key aspects of human intelligence have their roots in this social embedding, including (I argue): free will, emotion, empathy and self-modelling. To approach passing such a test it will be necessary for an intelligence to have these aspects, and hence will need to be socially embedded in the society which it will be tested in. The only effective way of doing this is by providing the entity with the necessary pre-programming or 'hooks' and then develop the intelligence with a considerable period of *in situ* training.

Another aspect of the Turing Test is the ability to "pass-off" as human. This would require the intelligence to be very close to human intelligence both in abilities and limitations. Given the wealth of contingent and context-dependent knowledge that is necessary to do this, this seems unlikely. However if such an intelligence was achieved it would compete directly with us, leaving us almost no area in which we could successfully dominate.

A modification of the Turing Test where the emphasis is on successful interaction rather than "passing-off" as human would be far more appropriate. It would be a much more feasible target than the full Turing Test and would result in intelligences that are far more pleasant to live with.

REFERENCES

Aydede, M. (1999). "Language of Thought Hypothesis: State of the Art."http://humanities.uchicago.edu/faculty/aydede/LOTH.SEP.html

Aydede, M. and G. Güzeldere (forthcoming). "Consciousness, Intentionality, and Intelligence: Some Foundational Issues for Artificial Intelligence." Journal of Experimental & Theoretical Artificial Intelligence.
Baron-Cohen, S., A. M. Leslie, et al. (1985). "Does the autistic child have a 'theory of mind'?" Cognition **21**: 37-46.
Baum, E. (1998). Manifesto for an Evolutionary Economics of Intelligence. Neural Networks and Machine Learning. C. M. Bishop. Berlin, Springer-Verlag**:** 285-344.
Bridgeman, B. (1992). "On the Evolution of Consciousness and Language." Pscoloquy **3**: http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?3.15.
Brooks, R. A. (1991). "Intelligence without Representation." Artificial Intelligence **47**(1-3): 139-159.
Burns, T. R. and E. Engdahl (1998). "The Social Construction of Consciousness Part 2:
Individual Selves, Self-Awareness, and Reflectivity." Journal of Consciousness Studies **2**: 166-184.
Byrne Richard, W. and A. Whiten (1988). Machiavellian intelligence : social expertise and the evolution of intellect in monkeys, apes and humans. Oxford ; New York, Clarendon Press : Oxford University Press.
Byrne Richard, W. and A. Whiten (1997). Machiavellian intelligence II : extensions and evaluations. Cambridge ; New York, NY, USA, Cambridge University Press.
Dautenhahn, K. (1997). "I Could Be You: The phenomenological dimension of social understanding." Cybernetics and Systems **28**(417-453).
Dennett Daniel, C. (1984). Elbow room : the varieties of free will worth wanting. Oxford, O.U.P.
Dennett, D. C. (1989). "The Origin of Selves." Cogito **3**(163-173).
Drescher, G. L. (1991). Made-up Minds - A Constructivist Approach to Artificial Intelligence. Cambridge, MA, MIT Press.
Edmonds, B. (1999). "Capturing social embeddedness: A constructivist approach." Adaptive Behavior **7**(3-4): 323-347.
Edmonds, B. (1999). "Gossip, Sexual Recombination and the El Farol Bar: modelling the emergence of heterogeneity." Journal of Artificial Societies and Social Simulation **2**(3).
Edmonds, B. (2000). "The Constructability of Artificial Intelligence (as defined by the Turing Test)." Journal of Logic Language and Information **9**: 419-424.
Edmonds, B. (2001). "Meta-Genetic Programming: Co-evolving the Operators of Variation." Elektrik **9**(1): 13-29.
Ernst, G. and A. Newell (1969). GPS: A Case Study in Generality and Problem Solving. New York, Academic Press.
French, R. M. (1990). "Subcognition and the limits of the Turing Test." Mind **99**: 53-65.
Goldberg, D. E., K. Deb, et al. (1989). "Messy genetic algorithms: Motiation, analysis, and first results." Complex Systems **3**: 493-530.
Gopnik, A. (1993). "How we know our minds: The illusion of first-person knowledge of
intentionality." Behavioural and Brain Sciences **16**: 1-14.
Hendriks-Jansen, H. (1997). "The Epistomology of Autism: Making a Case for an Embodied, Dynamic, and Historic Explanation." Cybernetics and Systems **28**: 359-416.
Hoffman, J. (1993). Vorhersage und Erkenntnis [Anticipation and Cognition]. Germany: Hogrefe, Goettingen.
Hofstadter, D. R. (1985). Analogies and Roles in Human and Machine Thinking. Metamagical Themas. New York, Basic Books.
Holland, J. H. (1992). Adaptation in Natural and Artificial Systems. Cambridge, MA, MIT Press.
Jannink, J. (1994). Cracking and Co-evolving randomList. Advances in Genetic Programming. K. E. Kinnear. Cambridge, MA, MIT Press.
Koza, J. R. (1992). Genetic Programming: On the Programming of Computers by Means of Natural Selection. Cambridge, MA, MIT Press.
Koza, J. R. (1994). Genetic Programming II: Automatic Discovery of Reusable Subprograms. CCambridge: MA, MIT Press.
Kummer, H., L. Daston, et al. (1997). The social intelligence hypothesis. Human by Nature: between biology and the social sciences. W. e. al. Hillsdale, NJ, Lawrence Erlbaum Associates**:** 157-17.
Minsky, M. (2002). "The Emotion Machine."http://web.media.mit.edu/~minsky
Perlis, D. (1997). "Consciousness as Self-Function." Journal of Consciousness Studies **4**: 509-525.
Rolls, E. T. (2000). "Precis of the brain and emotion." Behavioural and Brain Sciences **23**(2): 177-.
Sloman, A. (2002). How many separately evolved emotional beasties live within us?" Emotions in Humans and Artifacts. R. Trappl, P. Petta and P. S. Cambridge, MA, MIT Press.
Stolzmann, W. (1998). Anticipatory Classifier Systems. Genetic Programming, University of Wisconsin, Madison, Wisconsin, Morgan Kaufmann.
Tolman, E. C. (1932). Purposive behavior in animals and men. New York, Appleton.
Turing, A. M. (1936). "On Computable Numbers, with an application to the Entscheidungsproblem." Proc London math Soc **42**: 230-65.
Weizenbaum, J. (1976). Computer Power and Human Reason: From Judgement to Calculation, W.H.Freeman.
Wolpert, D. H. and W. G. Macready (1995). No free lunch theorems for search, Santa Fe.