

Developing Agents Who Can Relate To Us

– putting agents in *our* loop via situated self-creation

Bruce Edmonds
Centre for Policy Modelling
Manchester Metropolitan University
<http://www.cpm.mmu.ac.uk/~bruce>

1 Introduction

In this paper I do not directly consider the question of how to make artificial agents so that humans can relate to them, but more the reverse: *how to produce artificial agents so that they can relate to us*. However, this is directly relevant to human-computer interaction since we, as humans, are used to dealing with entities who can relate to us, so such an ability in agents could mark a shift away from merely using agents as *tools* towards forming *relationships* with them.

The basic idea is to put the human into the developmental loop of the agent so that the agent co-develops an identity that is intimately bound up with ours. This will give it a sound basis with which to base its dealings with us, enabling *its* perspective to be in harmony with our own in a way that would be impossible if one attempted to *design* such an empathetic sociality into it. The development of such an agent could be achieved by mimicking the early human development in important respects – i.e. by socially situating it within a human culture.

The implementation details that follow derive from a speculative theory of the development of the human self that will be described. This may well be wrong but it seems clear that something of this ilk does occur in the development of young humans (Werner 1999, Edmonds & Dautenhahn 1999). So the following can be seen as simply *a* method to enable agents to develop the required abilities – other methods and processes may have the same effect.

2 The inadequacy of the design stance for implementing a *deeper* sociality

I (amongst others) have argued elsewhere that if an agent is to be embedded in its society (which is necessary if it is to have a part in the social constructs) then one will not be able to *design* the agent first and *deploy* it in its social context second, but rather that a considerable period of *in situ* culturation will be necessary (Edmonds 1998). In addition to this it seems likely that several crucial aspects of the mind *itself* requires a society in order to develop, including *intelligence* (Edmonds Dautenhahn 1999, Edmonds 2001) and *free-will* (Edmonds 2000).

Thus rather than specify directly the requisite social facilities and mechanisms I take the approach of specifying the social “hooks” needed and then attempt to evolve the social skills within the target society. In this way key aspects of the agent develop already embedded in the society for which it is intended to deal with. In this way the agent can truly partake of the culture around it. This directly mirrors the way our intelligence is thought to have evolved (Kummer et al. 1997).

In particular I think that this process of embedding has to occur at an early stage of agent development for it to be most effective. In this paper I suggest that this needs to occur at an extremely basic stage: during the construction of the *self*. In this way the agent’s own self will

have been co-developed with its model of others and allow a deep empathy between agents and its society (in this case *us*).

3 A model of self construction

Firstly I outline a model of how the *self* may be constructed in humans. This model attempts to reconcile the following requirements:

- That the self is only experienced indirectly (Gopnik, 1993).
- That a self requires a strong form of self-reference (Perlis, 1997).
- That aspects of the self are socially constructed (Burns and Engdahl, 1998).
- “Recursive processing results from monitoring one’s own speech” (Bridgeman, 1992).
- That one has a “narrative centre” (Dennett, 1989).
- 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 model is to approach how we might provide the facilities for an agent to construct its *self* using social reflection via language use. Thus if the agent’s self 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 can be develop with its society 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 agent and returns decisions about new actions (that can include changing the focus of one’s perception and retrieving memories).
2. The agent does not have direct access to the workings of this basic process but only of its perceptions and actions, past and present.
3. This basic process seeks to model its environment and control it via its actions, including the other agents it can interact with. In particular is attempts to model the consequences of its actions (including speech acts).
4. This process naturally picks up and tries out selections of the communications it receives from these other agents and uses these as a basis (along with observed actions) for modelling the decisions of these agents.
5. As a result it becomes adapt at using communication acts to fulfil its own needs via others actions using its model of their decision making processes.
6. Using the language it produces itself it learns to model *itself* (i.e. to predict the decisions it will make) by applying its models of other agents 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 agent’s 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 are 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 experienced; the agent models others decision making using their utterances as fine-grained indications of their mental states (including intentions etc.); and finally that the agent models itself by applying its

model of others to itself (and vice versa). This seems to be broadly compatible with (Aydede and Güzeldere, forthcoming).

4 General consequences of this model of self construction

The important consequences of this model are:

- The fact that models of other agent and self-models are co-developed means that basic assumptions about self and other's cognition are similar. This can form the basis for true relationships.
- 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-like cognition and awareness with this language.
- Communication has several sorts of use: the first 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 such agents do not have access to the basic decision making processes they do have access to and can report on their linguistic 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

5 Towards implementing self-constructing agents

Working from the above model gives enough information to work towards an implementation. 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. 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 agent, and some upon the society it develops in. I will briefly outline a possibility for each.

The agent 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 et al., 2000) 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 to itself is not built-in. The language of communication needs to be a combinatorial one, one that can be combinatorially generated by the internal language and also deconstructed by the same.

The social situation of the agent needs to have a combination of complex cooperative and competitive pressures in it. The cooperation 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 et al., 1997). The complexity of the cooperative and competitive encourages the prediction of one's own decisions. A suitable environment is where, in order to gain substantial reward, cooperation is necessary, but that inter-group competition occurs as well as competition for the dividing up of the rewards that are gained by a cooperative group.

Many of the elements of this model have already been implemented in pilot systems (e.g. Drescher, 1991; Edmonds, 1999; Stoltzmann et al., 2000), but there is still much to be done.

6 Consequences for agent production and use

If we develop agents in this way, allowing them to learn their *selves* from within a human culture we may have developed agents such that we can relate to them because they will be able to relate to us etc. The sort of social games which involve second guessing, lying, posturing, etc. will be accessible to the agent due to the fundamental empathy that is possible between agent and human. Such an agent would not be an 'alien' but (like some of the humans we relate to) all the more unsettling for that.

To achieve this goal we will have to at least partially abandon the design stance and move more towards a stance of an *enabling* stance and accept the necessity of considerable culturation of our agents within our society much as we do with our children.

7 Conclusion

If we want to put artificial agents truly into the "human-loop" then they will need to be able to reciprocate our ability to relate to them, including relating to them relating to us etc. In order to do this it is likely that the development of the agent's self-modelling will have to be co-developed with its modelling of the humans it interacts with - just as our self-modelling have started to be influenced by our interaction with computers and robots (Turkle, 1983). One algorithm for this has been suggested which is backed up by a theory of the development of the human self. Others are possible.

References

- Aydede, M. (1999). Language of Thought Hypothesis: State of the Art. <<http://humanities.uchicago.edu/faculty/aydede/LOTH.SEP.html>>
- Aydede, M. and Güzeldere, G. (forthcoming). Consciousness, Intentionality, and Intelligence: Some Foundational Issues for Artificial Intelligence. *Journal of Experimental & Theoretical Artificial Intelligence*. <<http://humanities.uchicago.edu/faculty/aydede/JETAI.MA&GG.pdf>>
- Barlow, H. (1992). The Social Role of Consciousness - Commentary on Bridgeman on Consciousness. *Pscoloquy*, **3**(19) Consciousness (4). <<http://www.cogsci.soton.ac.uk/psyc-bin/newpsy?article=3.19>>
- Bridgeman, B. (1992). On the Evolution of Consciousness and Language, *Pscoloquy*, **3**(15) Consciousness (1). <<http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy?3.15>>
- Bridgeman, B. (1992). The Social Bootstrapping of Human Consciousness – Reply to Barlow on Bridgeman on Consciousness, *Pscoloquy*, **3**(20) Consciousness (5). <<http://www.cogsci.soton.ac.uk/psyc-bin/newpsy?article=3.20>>
- Burns, T. R. and Engdahl, E. (1998). The Social Construction of Consciousness Part 2: Individual Selves, Self-Awareness, and Reflectivity. *Journal of Consciousness Studies*, **2**:166- 184. <[http://www.imprint.co.uk/jcs_5_2.html#The social construction of consciousness: Part 2:>](http://www.imprint.co.uk/jcs_5_2.html#The%20social%20construction%20of%20consciousness%3A%20Part%202%3E) (abstract)
- Dennett, D. C. (1989) The Origin of Selves, *Cogito*, **3**:163-173. <<http://ase.tufts.edu/cogstud/papers/originss.htm>>
- Drescher, G. L. (1991). *Made-up Minds, a constructivist approach to artificial intelligence*. Cambridge, MA: MIT Press.
- Edmonds, B. (1998). Social Embeddedness and Agent Development. UKMAS'98, Manchester, December 1998. <<http://www.cpm.mmu.ac.uk/cpmrep46.html>>.
- Edmonds, B. (1999). Capturing Social Embeddedness: a Constructivist Approach. *Adaptive Behavior*, **7**(3/4), in press.
- Edmonds, B. (2000). Towards Implementing Free-Will. AISB2000 Symposium on *How to Design a Functioning Mind*, Birmingham, April 2000. <<http://www.cpm.mmu.ac.uk/cpmrep57.html>>
- Edmonds, B. (2001). The Constructability of Artificial Intelligence, *Journal of Logic, Language and Information*, in press. <<http://www.cpm.mmu.ac.uk/cpmrep53.html>>
- Edmonds, B. and Dautenhahn, K. (1998). The Contribution of Society to the Construction of Individual Intelligence. *Socially Situated Intelligence: a workshop held at SAB'98*, August 1998, Zürich. <<http://www.cpm.mmu.ac.uk:80/cpmrep42.html>>
- Hoffman, J. (1993). *Vorhersage und Erkenntnis [Anticipation and Cognition]*. Goettingen, Germany: Hogrefe.
- Gopnik, A. (1993) How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioural and Brain Sciences*, **16**:1-14.
- Koza, J. R. (1992). *Genetic Programming: the programming of computers by means of natural selection*. Cambridge, MA: MIT press.
- Kummer, H., Daston, L., Gigerenzer, G. and Silk, J. (1997). The social intelligence hypothesis. In Weingart et. al. (eds.), *Human by Nature: between biology and the social sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, 157-179.
- Perlis, D. (1997). Consciousness as Self-Function, *Journal of Consciousness Studies*, **4**: 509-525. <[http://www.imprint.co.uk/jcs_4_5-6.html#Consciousness as](http://www.imprint.co.uk/jcs_4_5-6.html#Consciousness%20as)> (abstract)
- Stolzmann, W., Butz, M. V., Hoffman, J. and Goldberg, D. E. (2000). First Cognitive Capabilities in the Anticipatory Classifier System. IlliGAL Report No. 2000008, Illinois Genetic Algorithms Laboratory, University of Illinois, Urbana, IL, USA. <<ftp://ftp-illigal.ge.uiuc.edu/pub/papers/IlliGALs/2000007.ps.Z>>
- Turkle, S. (1984). *The Second Self, computers and the human spirit*. London: Granada.
- Werner, E. (1999): The Ontogeny of the Social Self. Towards a Formal Computational Theory. In: Dautenhahn, K. (ed.) *Human Cognition and Social Agent Technology*, John Benjamins Publishing Company, 263-300.