Selection Pressures for a Theory-of-Mind Faculty in Artificial Agents

J. Noble, T. Hebbron, J. van der Horst, R. Mills, S. T. Powers, and R. A. Watson

Science and Engineering of Natural Systems group
School of Electronics and Computer Science
University of Southampton, UK
jn2@ecs.soton.ac.uk

Extended Abstract

To have a theory of mind (ToM) is to anticipate the behaviour of other agents by considering what they want and what they know. It requires a representation of the environment that includes the internal states (e.g., beliefs) of other agents. Adult humans generally possess a ToM ability, demonstrated by reasoning like "he did not see the chocolate being switched from the red box to the blue one, so I predict he will choose the red box." Note the distinction between what the speaker believes to be true, and what the speaker believes about the other agent's belief states. ToM is of interest in developmental psychology (when and how do children acquire it?) and primatology (do our near relatives possess it?).

In this project we ask: in an evolving population of social agents, under what circumstances would a ToM ability be selected for? Using simulation to identify the ecological niches that produce selection pressure for ToM should cast light on its origin in humans and on when we should expect to see it in other animals. We build on earlier work by Takano and Arita (2006).

To operationalize ToM we borrow a hierarchy of cognitive architectures from Dennett (1987). A zero-order intentional agent (often seen in ALife work) is purely reactive to its perceptual inputs. A first-order agent builds on this by including internal state that has a mapping relation with the environment, e.g., remembering where a predator was last seen. A second-order agent has basic ToM, i.e., it is equipped with a world-model that includes the internal states of other agents (e.g., "there's a predator behind that tree, but my friend hasn't seen it yet."). Third- and higher-order agents include a recursive aspect, i.e., a model of what I think he thinks I am thinking.

Low-order agents are logically prior, but the evolution of higher-order agents like ourselves is not inevitable. ALife and related work (Braitenberg, 1984) have shown that outwardly sophisticated behaviours can be produced by simple underlying mechanisms. The evolutionarily stable strategy will sometimes remain zero- or first-order and this will depend on aspects of the ecological niche, such as the nature of the payoff matrix for agent interactions and the degree of perceptual overlap between agents. We tested these ideas in simulation by constructing a range of different social environments and running invasion studies, in which a population of (n)-order agents is exposed to an infrequent (n+1)-order mutant. If the higher-order mutant is fitter and thus able to invade, this indicates selection pressure for more advanced ToM abilities.

Results confirm that fragmented perception (not all agents see the same things) and socially relevant payoff matrices (my payoff depends on both our actions) are necessary for ToM to evolve. More specifically, competitive rather than cooperative interactions produce greater selection pressure for ToM. This finding is a challenge for the common association between ToM and human language (Grice, 1969) as the latter requires a cooperative context. Something about the early human ecological niche must have combined cooperative and competitive contexts in a near-unique way.

References

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. MIT Press, Cambridge, MA.

Dennett, D. C. (1987). *The Intentional Stance*. MIT Press / Bradford Books, Cambridge, MA.

Grice, H. P. (1969). Utterer's meaning and intention. *Philosophical Review*, 68:147–177.

Takano, R. and Arita, T. (2006). Asymmetry between even and odd levels of recursion in a theory of mind. In Rocha, L. M., Yaeger, L. S., Bedau, M. A., Floreano, D., Goldstone, R. L., and Vespignani, A., editors, *Artificial Life X: Proceedings of the Tenth International Conference on Artificial Life*, pages 405–411. MIT Press, Cambridge, MA.