awareness magazine
SELF-AWARENESS IN AUTONOMIC SYSTEMS
10.2417/3201309.005095

The formal characterization of self-deception

Andrew J. I. Jones

Current work investigates applying formal logic to the definition of self-deception in humans, and highlights the relevance of the phenomenon to research in informatics.

Awareness is of central interest not only to cognitive scientists but also to computer scientists who are developing models of intelligent, adaptive systems. Self-awareness is a significant aspect of awareness, given that some, at least, of the agents in such systems are capable of representing and reasoning about both the cognitive processes of other agents and also their own. Any comprehensive attempt to analyse and characterize self-awareness will have to consider ways in which it may be constrained, among which, it has often been maintained, self-deception figures prominently, at least in humans.

The focus of our ongoing work is the application of logic—specifically modal logics of belief—to the task of formally characterizing a group of types of self-deception. The results so far indicate that it is possible to carry through that task without resorting to paraconsistent logic[1] while still making explicit the nature of the apparent inconsistency that self-deception exhibits. They also indicate that the type of belief logic commonly chosen in theoretical artificial intelligence is totally unsuited to this task because it contains strong awareness assumptions as axioms. The latter are usually referred to as the 'positive introspection axiom' (the modal schema 4, which in the case of belief says that if an agent believes that p then he believes that he believes that p), and the 'negative introspection axiom' (the modal schema 5, which in the case of belief says that if an agent does not believe that p then he believes that he does not believe that p).

Consequently, our work focuses on the development of a formal-logical conceptual model, using logics which can formally model the critical concepts, so that a small 'family' of types of self-deception can be clearly characterized. These include, for instance, the situation in which an agent does not believe that p, but believes that he does believe that p,[2] and the situation in which an agent believes that p, but believes that he does not believe that p. The approach taken reflects the methodological position described in a recently published paper,[3] in which we argue that—not least in the design of sociotechnical systems—the construction of computational models should be informed and guided by formal models of the key concepts involved, rather than by a merely informal, natural-language description of the proposed system. Thus we see the current work as aiming to provide a firm conceptual foundation for future computationally oriented system design.

Figure 1. *Michel de Montaigne (1533–1592). His remark about self-deception[2] was the point of departure for Jaakko Hintikka's formal-logical analysis in his book 'Knowledge and Belief.'[4] Our work criticizes and develops Hintikka's approach. (Image credit: © Georgios Kollidas, Dreamstime.com.)*

It might be contended that there would be no interest in designing an intelligent system that was itself capable of

Continued on next page

self-deception, except perhaps in the context of simulation studies in cognitive science. However, many intelligent systems would be expected to be able to represent, and to reason about, the belief states of the human agents with which they interact. And clearly their understanding of those human agents, and of the practical reasoning processes driving their actions, will be diminished if they—the artefacts—are incapable of accommodating self-deception.

Moreover, a recent book by the distinguished evolutionary biologist Robert Trivers[5] provides a fascinating new perspective on the importance of self-deception. While the traditional view among psychiatrists and psychologists has perhaps been to view self-deception as essentially a defence mechanism, Trivers assembles evidence from various sources suggesting that a distinct strategic (i.e., offensive in contrast to defensive) advantage may arise from the capacity to self-deceive: it enhances the ability to deceive others. Many computer scientists have long been interested in communicative deception, for the obvious reason that it is a significant feature of the way in which intelligent agents often interact in negotiation, and is a fundamental aspect of computer security.

If Trivers' central thesis is right, then the study of deception in communication among complex, reflective and adaptive systems should go hand in hand with the study of self-deception. Thus the expected impact of the current work is on the development of an enhanced understanding of strategic interaction in multiagent systems, and thereby on the future of cybersecurity. However, the next steps of our work in this area will seek to determine whether there are also other forms of self-deception that are not adequately covered by the proposed formal theory. We will also begin the investigation of how to describe the dynamics of self-deception, regarding the factors that are operative in generating and maintaining self-deceptive states.

Author Information

Andrew J. I. Jones
King's College London
London, United Kingdom

References

1. In contrast to the position advocated in N. C. A. da Costa and S. French, *Belief, contradiction and the logic of self-deception*, **Am. Philos. Quart. 27** (3), pp. 179–197, 1990.
2. Consider the remark of Montaigne: 'Some make the world believe that they believe what they do not believe. Others, in greater number, make themselves believe it.' D. M. Frame ed., **The Complete Works of Montaigne**, p. 322, Stanford University Press, CA, 1957.
3. A. J. I. Jones, A. Artikis, and J. V. Pitt, *The design of intelligent socio-technical systems*, **AI Rev. 39** (1), pp. 5–20, 2013.
4. J. Hintikka, **Knowledge and Belief: An Introduction to the Logic of the Two Notions**, pp. 124–125, Cornell University Press, Ithaca, NY, 1962.
5. R. Trivers, **Deceit and Self-Deception**, Allen Lane-Penguin Books, London, 2011.