# Self-deception and the logic of belief

Andrew Jones

King's College London, UK

The point of departure for this presentation is the brief discussion of self-deception that appears in Hintikka's book *Knowledge and Belief: An Introduction to the Logic of the Two Notions, Cornell UP, 1962*. Hintikka starts from a remark by Montaigne: "Some make the world believe that they believe what they do not believe; others, in greater number, make themselves believe it", and gives a formal treatment of (the second part of) Montaigne's remark that parallels Hintikka's analysis of Moore's puzzle about saying and disbelieving. Those analyses depend crucially on the 4. schema for the logic of belief (later dubbed the 'positive introspection schema'): $B_a p \to B_a B_a p$.

It will be argued that Montaigne's remark indicates just one of a small group of 'self-deception positions', the others of which are inconsistent if the logic of belief is that of a (relativised) modal system of type KD4 (Hintikka's choice), and all of which are inconsistent if KD45 is adopted (commonly the choice in AI).

The presentation will show how to characterise that group of 'self-deception positions' consistently using KD as the logic of belief, and provides an alternative treatment of Montaigne's remark and Moore's puzzle.

Why should researchers in Informatics concern themselves with self-deception? At least two reasons: first, there is already a good deal of interest in the phenomenon of awareness in Cognitive Science, and among those computer scientists who are developing models of self-organising, adaptive systems. Self-awareness, and thus also constrained self-awareness, of which self-deception is arguably an instance, is central to those interests. Secondly, the recent book by the distinguished evolutionary biologist Robert Trivers provides a fascinating new perspective on the importance of self-deception. While the traditional view among psychiatrists and psychologists has perhaps been that self-deception is 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 obvious reasons. If Trivers' central thesis is right, then the study of deception in communication among complex, reflective systems should perhaps go hand-in-hand with the study of self-deception.