South American Journal of Logic
Vol. 1, n. 2, pp. 387–400, 2015
ISSN: 2446-6719

On the Logic of Self-deception

Andrew J.I. Jones

Abstract

In his classic work on the logic of knowledge and belief, Jaakko Hintikka gave a brief analysis of the type of self-deception that is expressed in the following remark by Michel de Montaigne: "Some make the world believe that they believe what they do not believe; others, in greater number, make themselves believe it." Hintikka's account not only gives a logically consistent representation of this species of self-deception, but also explains the apparent incoherence of the self-deceiver's belief state.

This paper argues that Montaigne's remark describes just one of a small group of 'self-deception positions', the others of which cannot be consistently represented in the logic of belief used by Hintikka. It is shown how each member of the group of 'self-deception positions' can be characterized consistently, using a logic of belief weaker than Hintikka's. An alternative explanation is then offered of the incoherence latent in self-deception, and the account is extended to incorporate an analysis of Moore's puzzle about 'saying and disbelieving'.

1 Introduction

Consider the following example: a father does not believe that his daughter has been cured of her drug addiction, but nevertheless he manages to persuade himself that he does believe that she has been cured. This paper is premised on the conjecture that, while such conjunctions of beliefs are logically possible, they are nevertheless decidedly odd; so the challenge is two-fold: to show how these beliefs – which will here be assumed to be examples of self-deception – can be consistently characterized, and to explicate the nature of the oddity that they embody.

In (da Costa and French 1990) the authors argue that a formal account of the nature of self-deception calls for the use of paraconsistent logic. They aim to "...liberate discussion of self-deception from the shackles of a purely classical doxastic logic ..." (op. cit., p.179). Whilst they are not fully explicit about what they mean by 'purely classical doxastic logic', it seems clear from what they do say that a logic of belief

in which the concept is interpreted as a (relativized) normal modality in the sense of (Chellas 1980), and in which the D schema

D. $B_a p \rightarrow \neg B_a \neg p$

is adopted, would count for them as an instance. The D schema is obviously a consistency requirement; it will here be argued – contra da Costa and French – that a formal representation of the type of self-deception embodied in the father-daughter example can be consistently expressed in a logic of belief that accepts the D schema. In this regard, as will be seen, the approach will follow (Hintikka 1962), although it will also be shown why Hintikka's belief-logic is in other respects too strong to accommodate some other examples that closely resemble the father-daughter case.

A recent book by the distinguished evolutionary biologist Robert Trivers (Trivers 2011) 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.[1]

There are also reasons for researchers in Informatics to concern themselves with self-deception: 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.[2] Self-awareness, and thus also constrained self-awareness, of which self-deception is arguably an instance, is central to those interests. Secondly, 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 go hand-in-hand with the study of self-deception.

The paper proceeds as follows: section 2 introduces the passage from Montaigne that forms the point of departure for Hintikka's brief discussion of self-deception in (Hintikka 1962). Section 3 describes a method for generating a class of 'belief positions' relevant to the kind of example Montaigne described, using a logic of belief based on the modal system KD, and argues that Hintikka identified just one section of a broader family of 'self-deception positions', most of which turn out to be inconsistent in Hintikka's logic of belief because it includes the 'positive introspection' axiom, the modal 4 schema,

[1]So one is here tempted to insert a second example: consider a politician A who does not believe that state B possesses weapons of mass-destruction, but gets himself to believe that he does believe it, thereby enhancing his capacity to deceive others. A would be a self-deceiver, in contrast to A the liar, who does not believe that there are WOMDs in B, believes that he doesn't believe it, but tries to get others to believe that he does believe it.

[2]See, for instance, www.awareness-mag.eu for evidence of a considerable body of research in AI, Cognitive Science and Robotics on awareness and self-awareness.

in addition to the D schema. Section 4 then explains how Hintikka's analysis of the Montaigne example parallel's Hintikka's account of G.E. Moore's puzzle about 'saying and disbelieving', and indicates that both analyses rely on the modal 4 schema. Section 5 proposes an alternative account, compatible with the adoption of KD for the logic of belief. Section 6 concludes, and offers reasons for suspecting that the proposed account of self-deception does not provide an exhaustive taxonomy.

2 Montaigne's comment

At pages 124-125 of (Hintikka 1962) there is a brief discussion and analysis of a passage from Michel de Montaigne. Hintikka quotes the passage as follows:

> Some make the world believe that they believe what they do not believe. Others, in greater number, make themselves believe it. (Montaigne 1957, p.322)

Hintikka interprets the second sentence here as a statement to the effect that some agent believes that he believes something that in fact he does not believe. Accordingly, Hintikka's representation in his logic of belief is:

(1) $\neg B_a p \wedge B_a B_a p$

where a is an agent.

Hintikka's analysis of Montaigne's statement will be the subject of critical discussion below, in section 4. But what will here be accepted, at least initially, is that the belief-state to which Montaigne's second sentence refers may be given an interpretation that would be appropriately represented by (1), that (1) is a logically consistent[3] conjunction (as indeed it is in Hintikka's belief-logic), and that it is appropriate to describe (1) as representing a form of self-deception.

But if (1) represents a type of self-deception, then it would seem that

(2) $B_a p \wedge B_a \neg B_a p$

does so too. For if an agent can get the world (and himself) to believe that he believes what he does not believe, then surely he could get the world (and himself) to believe that he does not believe what he in fact believes. Similarly, perhaps,

(3) $B_a p \wedge B_a B_a \neg p$

[3]Hintikka used the term *defensible* rather than *consistent*, but the reasons for that difference of terminology are not the present concern.

might also be classified as a species of self-deception.

What is needed is a means of generating an overall picture of the class of those types of conjunctions of belief sentences that are exemplified by (1), (2) and (3) – that is, of generating the logical space within which the conjunctions that are relevant to the characterization of these sorts of self-deception can be identified.

To this end, the combinatory method of maxi-conjunctions, earlier developed by Kanger for classifying types of rights-relations in the spirit of Hohfeld, will prove to be useful; (see (Lindhal 2001) for an overview of Kanger's work on this topic, and for references to Kanger's original papers, and (Jones and Sergot 1992) for a further illustration of application of the method.)

3 Generating belief positions

For the logic of the belief modality, a (relativized) normal modality of type KD (according to the classification system presented in (Chellas 1980)) will be used. In essence, the choice of a normal modal logic means that the operator is closed under logical consequence[4]; in addition, since the logic is of type KD it also contains the axiom schema

D. $B_a p \rightarrow \neg B_a \neg p$

The D schema is of course equivalent to $\neg(B_a p \wedge B_a \neg p)$ and its adoption therefore means that a type of consistency constraint is imposed on agents' beliefs.

The procedure will consist of four steps, as follows:

(i) Using a belief-logic of type KD, first generate an exhaustive list of the possible 'B-positions', i.e., belief-sentences with a single belief operator of form $B_a p$, $B_a \neg p, \ldots$ and so on, or their negations.

(ii) Then generate an exhaustive list of the class of possible 'BB-positions', i.e., belief-sentences containing nested belief operators of form $B_a \beta, B_a \neg \beta \ldots$ and so on, or their negations, in which β is any one of the B-positions generated by step (i).

(iii) Then conjoin each of the B-positions with each of the BB-positions, to form a list of '$B\wedge$ BB-positions'.

(iv) Then, from that list, extract those positions that can plausibly be said to represent a type of self-deception.

[4]Closure under logical consequence requires that agents believe all of the logical consequences of that which they believe. This is of course an idealization, but a harmless one for the purposes of the present exercise.

Step(i): B-positions

Starting from $B_a p$ insert the negation sign in each of the available places to generate three further sentences $B_a \neg p, \neg B_a p, \neg B_a \neg p$. These four expressions may be displayed as the two truth-functional tautologies:

Bdis1 $B_a p \vee \neg B_a p$

Bdis2 $B_a \neg p \vee \neg B_a \neg p$

Obviously, just one disjunct in each of Bdis1 and Bdis2 must be true, for any proposition p and any agent a; there are four available combinations:

(B0) $B_a p \wedge B_a \neg p$

(B1) $B_a p \wedge \neg B_a \neg p$

(B2) $\neg B_a p \wedge B_a \neg p$

(B3) $\neg B_a p \wedge \neg B_a \neg p$

This list can be simplified. First, (B0) is removed because it does not represent a logically possible position: it is inconsistent with the D schema. In (B1), the second conjunct can be removed because – in virtue of the D schema – it is logically implied by the first conjunct. Similarly, the first conjunct of (B2) can be removed, since it is implied by the second conjunct. So the resulting revised list of B-positions is:

(B1) $B_a p$

(B2) $B_a \neg p$

(B3) $\neg B_a p \wedge \neg B_a \neg p$

It may readily be shown that, for any agent a and any proposition p, precisely one of (B1)-(B3) must hold.

Step(ii): BB-positions

First, prefix B_a to each of (B1), (B2) and (B3); then prefix $B_a \neg$ to each of (B1), (B2) and (B3); then prefix the negation sign to each of those six belief expressions, and display the resulting twelve expressions as six tautologies, as follows:

BBdis1 $B_a B_a p \vee \neg B_a B_a p$

BBdis2 $B_a\neg B_a p \vee \neg B_a \neg B_a p$

BBdis3 $B_a B_a \neg p \vee \neg B_a B_a \neg p$

BBdis4 $B_a \neg B_a \neg p \vee \neg B_a \neg B_a \neg p$

BBdis5 $B_a(\neg B_a p \wedge \neg B_a \neg p) \vee \neg B_a(\neg B_a p \wedge \neg B_a \neg p)$

BBdis6 $B_a \neg(\neg B_a p \wedge \neg B_a \neg p) \vee \neg B_a \neg(\neg B_a p \wedge \neg B_a \neg p)$

Obviously, just one disjunct in each of BBdis1-BBdis6 must be true, for any proposition p and any agent a; there are sixty-four available combinations. Of these sixty-four, it may be shown, by appeal to the properties of the logic, that just seven are consistent. Simplification of each of these seven conjunctions, to remove any conjunct that is logically implied by at least one other, results in the following list of BB-positions:

(BB1) $B_a B_a p$

(BB2) $B_a B_a \neg p$

(BB3) $B_a \neg B_a p \wedge B_a \neg B_a \neg p$

(BB4) $B_a \neg B_a p \wedge \neg B_a \neg B_a \neg p \wedge \neg B_a \neg(\neg B_a p \wedge \neg B_a \neg p)$

(BB5) $B_a \neg B_a \neg p \wedge \neg B_a \neg B_a p \wedge \neg B_a \neg(\neg B_a p \wedge \neg B_a \neg p)$

(BB6) $\neg B_a \neg B_a p \wedge \neg B_a \neg B_a \neg p \wedge B_a \neg(\neg B_a p \wedge \neg B_a \neg p)$

(BB7) $\neg B_a \neg B_a p \wedge \neg B_a \neg B_a \neg p \wedge \neg B_a \neg(\neg B_a p \wedge \neg B_a \neg p)$

Step(iii): $B\wedge$ BB-positions

We here list, in three groups of seven conjunctions, the twenty-one positions that are formed by adding, respectively, (B1), (B2) and (B3) to each of (BB1)–(BB7).

First, the seven (B1)/(BB) cases:

(B1)/(BB1) $B_a p \wedge B_a B_a p$

(B1)/(BB2) $B_a p \wedge B_a B_a \neg p$

(B1)/(BB3) $B_a p \wedge B_a \neg B_a p \wedge B_a \neg B_a \neg p$

(B1)/(BB4) $B_a p \wedge B_a \neg B_a p \wedge \neg B_a \neg B_a \neg p \wedge \neg B_a \neg(\neg B_a p \wedge \neg B_a \neg p)$

(B1)/(BB5) $B_a p \wedge B_a \neg B_a \neg p \wedge \neg B_a \neg B_a p \wedge \neg B_a \neg(\neg B_a p \wedge \neg B_a \neg p)$

(B1)/(BB6) $B_a p \wedge \neg B_a \neg B_a p \wedge \neg B_a \neg B_a \neg p \wedge B_a \neg(\neg B_a p \wedge \neg B_a \neg p)$

(B1)/(BB7) $B_a p \wedge \neg B_a \neg B_a p \wedge \neg B_a \neg B_a \neg p \wedge \neg B_a \neg (\neg B_a p \wedge \neg B_a \neg p)$

The seven (B2)/(BB) cases:

(B2)/(BB1) $B_a \neg p \wedge B_a B_a p$

(B2)/(BB2) $B_a \neg p \wedge B_a B_a \neg p$

(B2)/(BB3) $B_a \neg p \wedge B_a \neg B_a p \wedge B_a \neg B_a \neg p$

(B2)/(BB4) $B_a \neg p \wedge B_a \neg B_a p \wedge \neg B_a \neg B_a \neg p \wedge \neg B_a \neg (\neg B_a p \wedge \neg B_a \neg p)$

(B2)/(BB5) $B_a \neg p \wedge B_a \neg B_a \neg p \wedge \neg B_a \neg B_a p \wedge \neg B_a \neg (\neg B_a p \wedge \neg B_a \neg p)$

(B2)/(BB6) $B_a \neg p \wedge \neg B_a \neg B_a p \wedge \neg B_a \neg B_a \neg p \wedge B_a \neg (\neg B_a p \wedge \neg B_a \neg p)$

(B2)/(BB7) $B_a \neg p \wedge \neg B_a \neg B_a p \wedge \neg B_a \neg B_a \neg p \wedge \neg B_a \neg (\neg B_a p \wedge \neg B_a \neg p)$

The seven (B3)/(BB) cases:

(B3)/(BB1) $\neg B_a p \wedge \neg B_a \neg p \wedge B_a B_a p$

(B3)/(BB2) $\neg B_a p \wedge \neg B_a \neg p \wedge B_a B_a \neg p$

(B3)/(BB3) $\neg B_a p \wedge \neg B_a \neg p \wedge B_a \neg B_a p \wedge B_a \neg B_a \neg p$

(B3)/(BB4) $\neg B_a p \wedge \neg B_a \neg p \wedge B_a \neg B_a p \wedge \neg B_a \neg B_a \neg p \wedge \neg B_a \neg (\neg B_a p \wedge \neg B_a \neg p)$

(B3)/(BB5) $\neg B_a p \wedge \neg B_a \neg p \wedge B_a \neg B_a \neg p \wedge \neg B_a \neg B_a p \wedge \neg B_a \neg (\neg B_a p \wedge \neg B_a \neg p)$

(B3)/(BB6) $\neg B_a p \wedge \neg B_a \neg p \wedge \neg B_a \neg B_a p \wedge \neg B_a \neg B_a \neg p \wedge B_a \neg (\neg B_a p \wedge \neg B_a \neg p)$

(B3)/(BB7) $\neg B_a p \wedge \neg B_a \neg p \wedge \neg B_a \neg B_a p \wedge \neg B_a \neg B_a \neg p \wedge \neg B_a \neg (\neg B_a p \wedge \neg B_a \neg p)$

Step (iv): the self-deception positions

Can grounds be found for eliminating some of these twenty-one cases as not plausible instances of self-deception? Well, (B1)/(BB1) and (B2)/(BB2) both concern positions in which what the agent believes he believes matches what he believes. Similarly, in (B3)/(BB3) what the agent believes about what he fails to believe matches what he fails to believe. So these are clearly not examples of self-deception. Of the eighteen remaining cases, eight may be said to represent positions in which, to varying degrees, the agent just lacks full awareness of his belief state, rather than being in a state of self-deception. Those eight are: (B1)/(BB5), (B1)/(BB6), (B1)/(BB7), (B2)/(BB4), (B2)/(BB6), (B2)/(BB7), (B3)/(BB4), (B3)/(BB5). (The first six of those are cases where the agent fails to be fully aware of what he believes, and the last two are cases

where the agent fails to be fully aware of what he does not believe.) Finally, in case (B3)/(BB7), the agent has no positive beliefs regarding the truth or falsity of p, and no positive beliefs about whether or not he believes $p/\neg p$. So that too can scarcely be deemed to be a case of self-deception.

If those grounds for elimination are accepted, the nine cases that still remain are: (B1)/(BB2), (B2)/(BB1), (B1)/(BB3), (B1)/(BB4), (B2)/(BB3), (B2)/(BB5), (B3)/(BB1), (B3)/(BB2) and (B3)/(BB6). These are re-labelled below as (SD1)-(SD9), respectively. The positions (SD1)-(SD9) are mutually exclusive: at most one of them can hold for any given proposition p and agent a.

(SD1) $B_ap \wedge B_aB_a\neg p$

(SD2) $B_a\neg p \wedge B_aB_ap$

(SD3) $B_ap \wedge B_a\neg B_ap \wedge B_a\neg B_a\neg p$

(SD4) $B_ap \wedge B_a\neg B_ap \wedge \neg B_a\neg B_a\neg p \wedge \neg B_a\neg(\neg B_ap \wedge \neg Ba\neg p)$

(SD5) $B_a\neg p \wedge B_a\neg B_ap \wedge B_a\neg B_a\neg p$

(SD6) $B_a\neg p \wedge B_a\neg B_a\neg p \wedge \neg B_a\neg B_ap \wedge \neg B_a\neg(\neg B_ap \wedge \neg B_a\neg p)$

(SD7) $\neg B_ap \wedge \neg B_a\neg p \wedge B_aB_ap$

(SD8) $\neg B_ap \wedge \neg B_a\neg p \wedge B_aB_a\neg p$

(SD9) $\neg B_ap \wedge \neg B_a\neg p \wedge \neg B_a\neg B_ap \wedge \neg B_a\neg B_a\neg p \wedge B_a\neg(\neg B_ap \wedge \neg B_a\neg p)$

It will be observed that none of (SD1)-(SD9) corresponds exactly to (1) (above, p. 389), the initial attempt to represent the second part of Montaigne's remark. That is because the position-generation method employed forces consideration of whether, in addition to

(1) $\neg B_ap \wedge B_aB_ap$

it is also the case that $\neg B_a\neg p$. If it *is*, then Montaigne's remark is properly represented by (SD7); but if is *not*, then the appropriate representation is (SD2), the first conjunct of which logically implies the first conjunct of (SD7), in virtue of the D schema.

Similarly, none of (SD1)-(SD9) corresponds exactly to (2) (above, p.389). This is because the method employed forces consideration of whether, in addition to

(2) $B_ap \wedge B_a\neg B_ap$

it is also the case that $B_a\neg B_a\neg p$. If it *is*, then the appropriate representation is (SD3), but if it is *not*, then the appropriate form is (SD4). (The second and third conjuncts of (SD3) are together logically equivalent to $B_a(\neg B_a p \wedge \neg B_a\neg p)$ which, in virtue of the D schema, logically implies the fourth conjunct of (SD4).)

It is clear that (SD2) is just the result of replacing p by $\neg p$ in (SD1), and applying the property of the closure of the belief modality under logical equivalence. So (SD1) and (SD2) do not represent distinct types of self-deception: each represents the situation in which what a believes he believes is itself the denial of what he believes. Similarly, (SD3) and (SD5) represent one and the same type of self-deception; (SD4) and (SD6) represent one and the same type of self-deception; and (SD7) and (SD8) represent one and the same type of self-deception.

There remain then just five members of what may be called the 'Montaigne-family of types of self-deception':

(SD2) $B_a\neg p \wedge B_a B_a p$

(SD3) $B_a p \wedge B_a\neg B_a p \wedge B_a\neg B_a\neg p$

(SD4) $B_a p \wedge B_a\neg B_a p \wedge \neg B_a\neg B_a\neg p \wedge \neg B_a\neg(\neg B_a p \wedge \neg B_a\neg p)$

(SD7) $\neg B_a p \wedge \neg B_a\neg p \wedge B_a B_a p$

(SD9) $\neg B_a p \wedge \neg B_a\neg p \wedge \neg B_a\neg B_a p \wedge \neg B_a\neg B_a\neg p \wedge B_a\neg(\neg B_a p \wedge \neg B_a\neg p)$

4 Hintikka on Moore and Montaigne

Sections 4.5-4.7 of (Hintikka 1962) contain an exposition and analysis of Moore's puzzle about saying and disbelieving. (For relevant references, see (Hintikka 1962, p.64).) Since there are some significant similarities between Hintikka's respective treatments of Moore's puzzle and Montaigne's remark about self-deception, the former will be considered first. In essence, Moore's puzzle poses the following challenge: explain what is odd about the conjunction "It is raining but I do not believe that it is raining" in a way that is compatible with the (surely correct) intuition that the conjunction itself is not logically inconsistent. In terms of his own logic of belief, Hintikka provided a possible solution, by showing that although the conjunction

(4) $(p \wedge \neg B_a p)$

is consistent, and the sentence

(5) $B_b(p \wedge \neg B_a p)$

is consistent (where $a \neq b$), the sentence

(6) $B_a(p \wedge \neg B_a p)$

is not.

So the oddity identified by Moore's puzzle is explained by the fact that, although (4) is consistent, the agent referred to by 'a' cannot consistently believe (4) to be true. Hintikka interpreted the belief modality as a (relativised) normal modal operator of type KD4. From the semantical point of view, KD4 is characterized by means of a standard model in which the (relativised) accessibility relation is assigned the properties of seriality and transitivity. Both of those properties are involved in the proof of the unsatisfiability of (6), i.e., of the proof that (6) is false at all worlds in any serial and transitive standard model.

As stated above (p.389), Hintikka represents the Montaigne remark about self-deception as

(1) $\neg B_a p \wedge B_a B_a p$

Like (4), sentence (1) is a consistent conjunction in Hintikka's belief-logic. However, Hintikka also notes that Montaigne went on to make a supplementary remark, following the passage quoted on p.389 above, as follows: "... being unable to penetrate what it means to believe." This further insight of Montaigne, Hintikka suggests, is captured by the fact that the following sentence is not consistent in his belief-logic:

(7) $B_a(\neg B_a p \wedge B_a B_a p)$

As in the case of (6), above, the proof of the unsatisfiability of (7) requires appeal to both the seriality and transitivity of the accessibility relation, thus further revealing the parallel between Hintikka's respective formal analyses of the Moore puzzle and the Montaigne example.

Elegant though it may appear, Hintikka's approach runs into difficulties as soon as other putative examples of self-deception are considered, such as those exhibited by sentences (2) and (3). Returning to the characterization of the Montaigne-family of self-deception positions, it may readily be seen that each of (SD1)-(SD6), inclusive, is inconsistent if the belief logic is interpreted not as KD, but as KD4. From the axiomatic point of view, the difference between KD and KD4 is the addition of the so-called 'positive introspection' schema:

$$B_a p \rightarrow B_a B_a p$$

So, perhaps not surprisingly, positive introspection of this sort, when added to KD, appears to eliminate the possibility of self-deception of types (SD1)-(SD6).[5] As has

[5]As an aside, it is worth noting at this point that if KD4 were to be strengthened to KD45 by addition of the so-called 'negative introspection' schema:

$$\neg B_a p \rightarrow B_a \neg B_a p$$

just been observed, Hintikka's analysis of the challenges presented by Moore's puzzle, and by Montaigne's supplementary remark, relies on his adoption of transitivity for the accessibility relation. So the question that now needs to be considered is this: can an alternative analysis of those two challenges be provided that is compatible with the adoption of KD for the logic of belief, and thus with the consistency of each of the members of the Montaigne-family of self-deception positions? The next section provides an affirmative answer.

5 An alternative account of Moore's puzzle

When an agent makes an assertion, it is ordinarily possible for any one of the four statements (a)-(d), about his communicative act and its content, to be true:

> (a) The agent believes what he is saying, and his assertion is reliable, in that its content is true.
>
> (b) The agent does not believe what he is saying, and the content of his assertion is not true.
>
> (c) The agent believes what he is saying, but he is mistaken, in as much as the content of his assertion is not true.
>
> (d) The agent does not believe what he is saying, but – unbeknown to him – it happens that the content of his assertion is true.

The oddity of the sentence in the Moore puzzle, "It is raining, but I do not believe that it is raining", is that the first of those four possibilities is eliminated – at least, it is eliminated if the logic of belief is assumed to be that of a normal modality, i.e., at least of type K. It may readily be shown that the conjunction

> (8) $B_a(p \wedge \neg B_a p) \wedge p \wedge \neg B_a p$

is logically inconsistent, if the belief modality is normal, even though (unless one uses a logic of belief as strong as Hintikka's) the first conjunct is consistent and, as has already been noted, the last two conjuncts themselves form a consistent conjunction.[6]

---

then *none* of (SD1)-(SD9) would represent a logically possible belief position. KD45 has frequently been the belief-logic of choice in Artificial Intelligence; even if there are good reasons for supposing that there would be no interest in designing an artefact that could itself exhibit self-deception, it is nevertheless the case that if future machines are to be able to interact effectively with human beings, as their 'artificial companions', then those machines should be equipped with a means of representing and reasoning about the belief-states of self-deceivers.

[6]In an earlier version of the argument (Jones and Kimbrough 2008, p.224), the inconsistency was expressed in terms of the beliefs of some agent b, not identical to a, who hears a's assertion of the

Thus the oddity of the Moore example lies in the fact that if the agent says "It is raining, but I do not believe that it is raining" and believes what he is saying, then what he is saying is false; conversely, if what he is saying is true, then he cannot believe that it is. Note also that the first conjunct of (8) is equivalent to

(8') $B_a p \wedge B_a \neg B_a p$

and so counts as a type of self-deception. According to the analysis offered above, it is of type (SD3) or (SD4), depending on whether or not $B_a \neg B_a \neg p$ is also the case. So it constitutes self-deception of a type that Hintikka's KD4 logic of belief fails to capture.

In a comparable fashion, the challenge conveyed in Montaigne's supplementary remark may be explained by noting that even though (contra Hintikka) sentence (7) is KD-consistent, the conjunction of (7) and (1) – i.e., (9) below – is KD-inconsistent:

(9) $B_a(\neg B_a p \wedge B_a B_a p) \wedge \neg B_a p \wedge B_a B_a p$

In other words, were a to assert "I believe that I believe that p, but do not in fact believe that p", his assertion could be reliable only if he does not believe it to be true.

Similarly, where (SDn) denotes any of (SD1)-(SD9), it may be shown that $B_a(\mathrm{SD}n)$ is KD-consistent but that the conjunction $B_a(\mathrm{SD}n) \wedge (\mathrm{SD}n)$ is KD-inconsistent. Thus the challenge embodied in Montaigne's supplementary remark is also met for each of the nine members of the Montaigne-family of self-deception positions.[7]

Moore sentence. The sentence

$$B_b(B_a(p \wedge \neg B_a p) \wedge p \wedge \neg B_a p)$$

is KD-inconsistent, but not K-inconsistent, although even in system K it may be shown that if b has the above belief then he believes q, where q is any proposition whatsoever. The informally stated explanation of Moore's puzzle offered by (Searle 1969, p.65, fn. 1) is like Hintikka's in that its focus is on the beliefs of the speaker, and does not make explicit the tension between the speaker's belief and reliability-of-content, which the present account – first published in (Jones 1983) – sees as the key to Moore's puzzle. A similar approach was later adopted by (Hilpinen 2002, pp. 83-84) in his analysis of a case of self-deception.

[7]At this point it is appropriate to indicate another point of difference between the account here offered of Moore's puzzle and the earlier accounts I published in (Jones 1983), (Jones and Kimbrough 2008) (and elsewhere). In the earlier versions, the agent's 'believing that what he is saying is true' was referred to as the agent's *sincerity*. However, following some interesting comments made by one of the reviewers of the present paper, I accept that there is something decidedly odd in describing the utterer of the Moore sentence as *sincere* if he believes that what he is saying is true – a self-deceptive belief, on my account! Indeed, on Hintikka's account, as we have seen, the agent logically cannot be *sincere* in that sense of the term: he cannot believe that what he is saying is true. But one may question whether *mere belief* in the truth of what one is saying is sufficient to guarantee the *sincerity* of one's utterance; perhaps one should require not only that the utterer believes what he is saying, but also that he is fully cognizant of that fact – as would ordinarily be the case for an intentional utterance. It may readily be shown that the conjunction

(10) $B_a(p \wedge \neg B_a p) \wedge K_a B_a(p \wedge \neg B_a p)$

6 Concluding remarks

While the paper has offered a formal characterization of a class of cases of self-deception akin to those alluded to by Montaigne, no claim is made to the effect that this account is exhaustive. Consider the case of a woman who demands constant care and attention from her husband on the grounds that she is unwell; maybe she is in fact ill, but maybe the *real* reason why she demands that the dominant profile of her relationship with her husband shall be that of carer-and-patient is not that she is ill, but that she is emotionally incapable of entering into a fully reciprocal loving relationship with him. In a similar vein, consider the case of a vegetarian who claims that his vegetarianism is grounded on such moral considerations as the rights of animals, when in fact the real reason behind his behaviour derives from a need to supress his cannibalistic desires.[8]

Can examples of these kinds be accommodated within the formal framework proposed for the Montaigne family? Should one say of the woman in the first case that she does not in fact believe that the reason for her behaviour is her illness, but that she has persuaded herself that she does believe it to be the reason? Or should one say, rather, that this is not a case of self-deception at all: she genuinely believes that her illness is the reason for her behaviour but her belief is mistaken, because – unbeknown to her – the real cause of her behaviour is quite different from the alleged reason? Or should one say that the woman is lying, attempting to get her husband to believe that the reason for her behaviour is her illness, even though she is fully aware that it is not in fact the reason? In the absence of further evidence, it is surely impossible to decide which description applies. But the example does at least indicate that a more comprehensive account of the Montaigne-family of self-deception positions should examine its relationship to a broad formal theory of types of mistaken or dishonest beliefs. Perhaps it also shows that a full-blown characterization of self-deception will need to be placed within a formal theory of practical reasoning, in which it is possible to compare an agent's beliefs about the reasons for his behaviour with the factors causing it. Those are matters for further investigation.

References

Chellas, B.F. (1980) *Modal Logic – an Introduction.* Cambridge University Press, Cambridge.

da Costa, N.C.A. and French, F. (1990) "Belief, Contradiction and the Logic of Self-Deception", *American Philosophical Quarterly*, Volume 27, no. 3.

is KD-inconsistent, if it is assumed (as is usual) that knowledge implies belief. So, on my account too, the utterer of the Moore sentence cannot be sincere, if *sincerity* is given this stronger interpretation.

[8]An example of this kind is given in (Erwin 1988, p.230).

Erwin, E. (1988) "Psychoanalysis and Self-Deception", in B. McLaughlin and A. Oksenberg Rorty, eds., *Perspectives on Self-Deception*, University of California Press, Berkeley, Los Angeles and London.

Hilpinen, R. (2002) "Some Remarks on Self-Deception: Mele, Moore and Lakatos", *Florida Philosophical Review*, Volume II, Issue 1.

Hintikka, J. (1962) *Knowledge and Belief – an Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca and London, 1962

Jones, A.J.I. (1983) *Communication and Meaning – an essay in applied modal logic*, Synthese Library vol. 168, D. Reidel, Dordrecht, Holland.

Jones, A.J.I. and Kimbrough, S.O. (2008) "The Normative Aspect of Signalling and the Distinction between Performative and Constative", *Journal of Applied Logic*, vol. 6(2), pp. 218-228.

Jones, A.J.I. and Sergot, M.J. (1992) "Formal specification of security requirements using the theory of normative positions", in Y. Deswarte et al., eds., *Computer Security-ESORICS 92, Proc. of the 2nd European Symposium on Research in Computer Security*, Springer Lecture Notes in Computer Science, vol.648, Springer-Verlag, pp. 103-121.

Lindahl, L. (2002) "Stig Kanger's Theory of Rights", in Holmström-Hintikka, G., Lindström, S., and Sliwinski, R., eds., *Collected Papers of Stig Kanger with Essays on his Life and Work*, Vol. II, Dordrecht, Holland: Kluwer Academic Publishers, pp. 151-171, 2001.

Montaigne, M. (1957) *The Complete Works of Montaigne*, Frame, D.M., ed., Stanford, California, 1957.

Searle, J.R. (1969) *Speech Acts – An Essay in the Philosophy of Language*, Cambridge UP, Cambridge.

Trivers, R. (2011) *Deceit and Self-Deception*, Allen Lane – Penguin Books, London.

Andrew J.I. Jones
Emeritus Professor, Department of Informatics
King's College
London, Strand, LONDON WC2R 2LS, UK
E-mail: andrewji.jones@kcl.ac.uk