The Logic of Self-Deception

Andrew J I Jones
Emeritus Professor, King's College London
Academic (Research) Visitor, Imperial College London
andrew.jones3@imperial.ac.uk
andrewji.jones@kcl.ac.uk

Self-Deception (intro)

Principal aim

- To show (in response to da Costa & French 90) that there are notions of self-deception that can be consistently characterised using a normal modality, with the D schema, for the logic of belief.
- D. $B_a p \rightarrow \neg B_a \neg p$
- *Equivalent to* $\neg (B_a p \ \& \ B_a \neg p)$

Self-Deception (intro)

- Should research in Informatics be concerned with self-deception ?
- Interest in *awareness* in Cognitive Science and among computer scientists 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.

Self-Deception (intro)

- 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.

Montaigne’s example

- Hintikka quotes Montaigne as follows:
- *Some make the world believe that they believe what they do not believe. Others, in greater number, make themselves believe it.*
- The second sentence implies that some agents believe that they believe something that in fact they do not believe.

Hintikka on Montaigne

- In his own logic of belief, Hintikka represented the situation Montaigne is alluding to in terms of the conjunction

$$(1) \quad \neg B_a p \ \& \ B_a B_a p$$

- *Accepted here*:

✓ (1) is a way of representing the belief-state described in Montaigne’s second sentence, and

✓ (1) is a logically consistent conjunction (as indeed it is in Hintikka’s belief-logic), and

✓ (1) represents a form of self-deception.

Other cases of Self-Deception ?

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

$$(2)\quad B_a p \;\;\&\;\; B_a \neg B_a p$$

does so too. 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

$$(3)\quad B_a p \;\;\&\;\; B_a B_a \neg p$$

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

Generating belief ‘positions’ (i)

- We need 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 we can identify the conjunctions that are relevant to the characterising these sorts of self-deception.
- To that 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.

Generating belief 'positions' (ii)

For the logic of the belief modality, use a (relativised) normal modality of type KD. Proceed as follows:

- Generate an exhaustive list of the possible 'B-positions', i.e., belief-sentences with a single belief operator of form $B_a p$, $B_a \neg p$,…….and so on, and their negations.
- Generate an exhaustive list of the class of possible 'BB-positions', i.e., belief-sentences with a nested pair of belief operators of form $B_a B_a p$, $B_a \neg B_a p$,……..and so on, and their negations.
- Conjoin each of the B-positions with each of the BB-positions, to form a list of 'B & BB-positions'.
- From that list extract those positions that can plausibly be said to represent a type of self-deception.

Generating belief ‘positions’ (iii)

B-positions

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

$$\text{Bdis1} \quad B_a p \lor \neg B_a p$$

$$\text{Bdis2} \quad B_a\neg p \lor \neg B_a\neg p$$

- Obviously, just one disjunct in each of Bdis1 and Bdis2 must be true; there are four available combinations:

Generating belief ‘positions’ (iv)

(B0) $B_a p$ & $B_a \neg p$ *(inconsistent because of D. schema)*

(B1) $B_a p$ & $\neg B_a \neg p$ *(simplifies to $B_a p$ because of D. schema)*

(B2) $\neg B_a p$ & $B_a \neg p$ *(simplifies to $B_a \neg p$ because of D. schema)*

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

So the B-positions are:

(B1) $B_a p$

(B2) $B_a \neg p$

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

Generating belief ‘positions’ (v)

BB-positions

- Starting from B_aB_ap insert the negation sign in each of the available places to generate seven further sentences $B_aB_a\neg p$, $B_a\neg B_ap$, $B_a\neg B_a\neg p$, $\neg B_aB_ap$, $\neg B_aB_a\neg p$, $\neg B_a\neg B_ap$, $\neg B_a\neg B_a\neg p$. Display as the four truth-functional tautologies:

BBdis1 $B_aB_ap \vee \neg B_aB_ap$

BBdis2 $B_a\neg B_ap \vee \neg B_a\neg B_ap$

BBdis3 $B_aB_a\neg p \vee \neg B_aB_a\neg p$

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

Generating belief ‘positions’ (vi)

- Obviously, just one disjunct in each of BBdis1, BBdis2, BBdis3 and BBdis4 must be true; there are sixteen available combinations. Of these sixteen, it may readily be shown, by appeal to the properties of the logic, that ten are inconsistent. The remaining six may be simplified, to generate the following list of BB-positions:

(BB1) B_aB_ap

(BB2) $B_aB_a\neg p$

(BB3) $B_a\neg B_ap$ & $B_a\neg B_a\neg p$

(BB4) $B_a\neg B_ap$ & $\neg B_aB_a\neg p$ & $\neg B_a\neg B_a\neg p$

(BB5) $B_a\neg B_a\neg p$ & $\neg B_aB_ap$ & $\neg B_a\neg B_ap$

(BB6) $\neg B_a\neg B_ap$ & $\neg B_a\neg B_a\neg p$

Generating belief ‘positions’ (vii)

B&BB-positions

- List, in three groups of six conjunctions, the eighteen positions that are formed by adding, respectively, (B1), (B2) and (B3) to each of (BB1)-(BB6).
- First, the six (B1)/(BB) cases:

(B1)/(BB1) $B_a p \;\&\; B_a B_a p$

(B1)/(BB2) $B_a p \;\&\; B_a B_a \neg p$

(B1)/(BB3) $B_a p \;\&\; B_a \neg B_a p \;\&\; B_a \neg B_a \neg p$

(B1)/(BB4) $B_a p \;\&\; B_a \neg B_a p \;\&\; \neg B_a B_a \neg p \;\&\; \neg B_a \neg B_a \neg p$

(B1)/(BB5) $B_a p \;\&\; B_a \neg B_a \neg p \;\&\; \neg B_a B_a p \;\&\; \neg B_a \neg B_a p$

(B1)/(BB6) $B_a p \;\&\; \neg B_a \neg B_a p \;\&\; \neg B_a \neg B_a \neg p$

Generating belief ‘positions’ (viii)

- The six (B2)/(BB) cases:

(B2)/(BB1) $B_a\neg p \;\&\; B_aB_ap$

(B2)/(BB2) $B_a\neg p \;\&\; B_aB_a\neg p$

(B2)/(BB3) $B_a\neg p \;\&\; B_a\neg B_ap \;\&\; B_a\neg B_a\neg p$

(B2)/(BB4) $B_a\neg p \;\&\; B_a\neg B_ap \;\&\; \neg B_aB_a\neg p \;\&\; \neg B_a\neg B_a\neg p$

(B2)/(BB5) $B_a\neg p \;\&\; B_a\neg B_a\neg p \;\&\; \neg B_aB_ap \;\&\; \neg B_a\neg B_ap$

(B2)/(BB6) $B_a\neg p \;\&\; \neg B_a\neg B_ap \;\&\; \neg B_a\neg B_a\neg p$

Generating belief ‘positions’ (ix)

- The six (B3)/(BB) cases:

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

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

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

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

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

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

Generating belief ‘positions’ (x)

- We can eliminate some of these eighteen cases as *not* plausible instances of self-deception.
- (B1)/(BB1) and (B2)/(BB2) concern positions in which what the agent believes he believes matches what he believes.
- 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.

Generating belief ‘positions’ (xi)

- Of the fifteen remaining cases, seven may be said to represent positions in which the agent just lacks full awareness of his belief state, rather than being in a state of self-deception: (B1)/(BB5), (B1)/(BB6), (B2)/(BB4), (B2)/(BB6), (B3)/(BB4), (B3)/(BB5), (B3)/(BB6).
- The first four of those are cases where the agent fails to be aware of what he believes, and the last three are cases where the agent fails to be aware of what he does not believe.

Self-Deception positions (i)

- Eight cases remain: (B1)/(BB2), (B2)/(BB1), (B1)/(BB3), (B1)/(BB4), (B2)/(BB3), (B2)/(BB5), (B3)/(BB1), (B3)/(BB2).
- Re-label these (SD1)-(SD8) :

(SD1) $B_a p \ \& \ B_a B_a \neg p$

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

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

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

(SD5) $B_a \neg p \ \& \ B_a \neg B_a p \ \& \ B_a \neg B_a \neg p$

(SD6) $B_a \neg p \ \& \ B_a \neg B_a \neg p \ \& \ \neg B_a B_a p \ \& \ \neg B_a \neg B_a p$

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

(SD8) $\neg B_a p \ \& \ \neg B_a \neg p \ \& \ B_a B_a \neg p$

Self-deception positions (ii)

- (SD2) is just the result of replacing *p* by ¬*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) & (SD5) represent one and the same type of self-deception; as do (SD4) & (SD6) and (SD7) & (SD8).

Self-Deception positions (iii)

- There remain just four members of what we shall call the ‘Montaigne-family of types of self-deception’:
  (SD2) $B_a \neg p \ \& \ B_a B_a p$
  (SD3) $B_a p \ \& \ B_a \neg B_a p \ \& \ B_a \neg B_a \neg p$
  (SD4) $B_a p \ \& \ B_a \neg B_a p \ \& \ \neg B_a B_a \neg p \ \& \ \neg B_a \neg B_a \neg p$
  (SD7) $\neg B_a p \ \& \ \neg B_a \neg p \ \& \ B_a B_a p$

Self-Deception positions (iv)

- None of (SD1)-(SD8) corresponds exactly to

(1) $\neg B_a p$ & $B_a B_a p$ (our first attempt to represent the second part of Montaigne's remark).

- the position-generation method forces us to consider

whether, in addition to (1), it's also the case that $\neg B_a \neg p$.

- If it *is*, then Montaigne's remark is represented by

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

- but if is *not*, the appropriate representation is

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

Self-Deception positions (v)

- Similarly, none of (SD1)-(SD8) corresponds exactly to

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

- The position-generation method forces us to consider whether, in addition to (2), it's also the case that $B_a \neg B_a \neg p$.
- If it *is*, then we have

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

- But if it is not, we have

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

Equivalently: $B_a p \ \& \ B_a \neg B_a p \ \& \ \neg(B_a B_a \neg p \ \text{v} \ B_a \neg B_a \neg p)$

Hintikka on Moore and Montaigne (i)

- 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 compatible with the (surely correct) intuition that the conjunction itself is not logically inconsistent.
- In terms of his own logic of belief (KD4), Hintikka provided a possible solution, by showing that although the conjunction

  (4) $p \mathbin{\&} \neg B_a p$ is consistent, and the sentence

  (5) $B_b(p \mathbin{\&} \neg B_a p)$ is consistent (where $a \neq b$), the sentence

  (6) $B_a(p \mathbin{\&} \neg B_a p)$ is not. So, although (4) is consistent, the agent referred to by '*a*' cannot consistently believe (4).

Hintikka on Moore and Montaigne (ii)

- Hintikka represents the Montaigne remark as

  (1) $\neg B_a p \;\&\; B_a B_a p$

- Like Moore’s conjunction, sentence (1) is consistent in Hintikka’s belief-logic.
- Hintikka notes that Montaigne went on to make a supplementary remark, as follows: “…..being unable to penetrate what it means to believe.” This further point from Montaigne, Hintikka suggests, is captured by the fact that the following sentence is *not* consistent in his belief-logic:

  (7) $B_a(\neg B_a p \;\&\; B_a B_a p)$

Hintikka on Moore and Montaigne (iii)

- As in the case of

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

the proof of the inconsistency of

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

requires appeal to both the D-schema

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

And the so-called ‘positive introspection’ schema

PI. $B_a p \ \rightarrow \ B_a B_a p$ (Hintikka adopts PI; I do not.)

(Note parallel between Hintikka’s respective formal analyses of the Moore puzzle and the Montaigne example.)

Hintikka on Moore and Montaigne (iv)

- Hintikka's approach runs into difficulties as soon as we consider other examples of self-deception.
- It may readily be seen that each of (SD2), (SD3) and (SD4) is *inconsistent* if the belief logic is interpreted not as KD (my choice), but as KD4 (Hintikka's choice). From the axiomatic point of view, the difference between KD and KD4 is the addition of the so-called 'positive introspection' schema

  PI. $B_a p \rightarrow B_a B_a p$
- So, perhaps not surprisingly, positive introspection eliminates the possibility of self-deception corresponding to the first three types in our Montaigne-family.

Hintikka on Moore and Montaigne (v)

- Self-deception as expressed by the fourth type in our Montaigne-family (SD7) would be inconsistent were the ‘negative introspection’ schema also to be adopted

  NI. $\neg B_a p \rightarrow B_a \neg B_a p$

- Consequently, in the logic of belief KD45 – which has quite standardly been the doxastic logic of choice in AI – *none* of the members of the Montaigne-family is a logically consistent expression.
- Jones, A.J.I., Artikis, A., Pitt, J.V., “The design of intelligent socio-technical systems”, *Artificial Intelligence Review* (2013), 39:5-20.

An alternative analysis (i)

- When an agent makes an assertion, it is ordinarily possible for any one of the following four statements about his communicative act and its content to be true:
- The agent is sincere, in that he believes that what he is saying is true, and his assertion is reliable, in that its content is true;
- The agent is insincere, and his assertion is not reliable (its content is untrue);
- The agent is sincere but mistaken, in as much as the content of his assertion is untrue;
- The agent is insincere, but – unbeknown to him – it happens that the content of his assertion is true.

An alternative analysis (ii)

- 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) $B_a(p \mathbin{\&} \neg B_a p) \mathbin{\&} p \mathbin{\&} \neg B_a p$

  is logically inconsistent, if the belief modality is normal; as already noted, the last two conjuncts themselves form a consistent conjunction.
- Note that the first conjunct is KD-consistent.

An alternative analysis (iii)

- The challenge conveyed in Montaigne’s supplementary remark may be similarly be explained since even though

  (7) $B_a(\neg B_a p \ \& \ B_a B_a p)$ is KD-consistent, the conjunction of

  (7) and (1): $B_a(\neg B_a p \ \& \ B_a B_a p) \ \& \ \neg B_a p \ \& \ B_a B_a p$

  is KD-inconsistent.

- So, were *a* to assert “I believe that I believe that *p*, but do not in fact believe that *p*”, his assertion logically could not be both sincere and reliable. Alternatively: if *a* believes, of himself, that he is in the belief-position represented by (1), then he cannot in fact be in that position; thus, if he is in belief-position (1) then he cannot believe that he is.

An alternative analysis (iv)

- Similarly, where (SD*n*) denotes any of (SD2), (SD3), (SD4), (SD7), it may be shown that B_a(SD*n*) is KD-consistent but that the conjunction

  B_a(SD*n*) & (SD*n*)

  is KD-inconsistent.
- Thus the challenge embodied in Montaigne’s supplementary remark is also met for each of the members of the Montaigne-family of types of self-deception.

Some references

- da Costa, N.C.A. & French, F.: “Belief, Contradiction and the Logic of Self-Deception”, *American Philosophical Quarterly*, Volume 27, no. 3, 1990.
- Hintikka, J.: *Knowledge and Belief – an Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca and London, 1962.
- Jones, A.J.I. & Sergot, M.J.: "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, 1992.
- Trivers, R.: *Deceit and Self-Deception*, Allen Lane – Penguin Books, London, 2011.