

Self-deception, intentions and contradictory beliefs

José Luis Bermúdez
 Department of Philosophy
 University of Stirling
 Stirling FK9 4LA
 Scotland
 jb10@stir.ac.uk

Philosophical discussion of self-deception aims to provide a conceptual model for making sense of a puzzling but common mental phenomenon. The models proposed fall into two groups – the intentionalist and the anti-intentionalist. Broadly speaking, intentionalist approaches to self-deception analyse the phenomenon on the model of other-deception - what happens in self-deception is parallel to what happens when one person deceives another, except that deceiver and deceived are the same person. Anti-intentionalist approaches, in contrast, stress what they take to be the deep conceptual problems involved in trying to assimilate self-deception to other-deception. Many of the arguments appealed to by anti-intentionalists suffer from failing to specify clearly enough what the intentionalist is actually proposing. In this paper I distinguish three different descriptions that an intentionalist might give of an episode of self-deception. Separating out the different possible ways in which an intentionalist might characterise an episode of self-deception allows us to evaluate some of the commoner objections that are levelled against the intentionalist approach. I end by offering a tentative argument in favour of intentionalism, suggesting that only intentionalist accounts of self-deception look as if they can deal with the selective nature of self-deception.

I am concerned here only with self-deception issuing in the acquisition of beliefs, putting to one side both cases of self-deception in which what is acquired is a propositional attitude that is not a belief (such as a hope or a fear), and cases in which the self-deception generates the retention rather than the acquisition of a belief. In general terms the intentionalist view is that a subject forms a self-deceiving belief that p when they intentionally bring it about that they believe that p . But there are, of course, different ways in which this might happen. Consider the following three situations.

(1) S believes that $\sim p$ but intends to bring it about that he acquires the false belief that

(2) S believes that $\sim p$ but intends to bring it about that he acquires the belief that p.

(3) S intends to bring it about that he acquires the belief that p.

There are three dimensions of variation here. First, there is the question of whether the subject actually holds the belief that p to be false. Second, there is the question of whether the subject actually intends to bring it about that he believes a falsehood. Third, there is the question of whether there is a belief such that the subject actually intends to bring it about that he acquires that belief. We can accordingly identify the following three claims, in ascending order of strength, that an intentionalist might make about the phenomenon of self-deceiving belief acquisition:

(A) A given episode of self-deception can involve intending to bring it about that one acquires a certain belief.

(B) A given episode of self-deception can involve holding a belief to be false and yet intending to bring it about that one acquires that belief.

(C) A given episode of self-deception can involve intending to bring it about that one acquires a false belief.

It is clear that holding (A) to be true is a bottom-level requirement on any intentionalist account of self-deception.

But (A) does not entail either (B) or (C). There is something puzzling about the idea that one might hold it to be the case that p and intend to bring it about that one acquires the belief that p. If one holds it to be the case that p then one already believes that p - and any intention to bring it about that one believe that p would be pointless. But one might, of course, have no views whatsoever about whether or not it is the case that p and yet intend to bring oneself to believe that p. This would be the position of a complete agnostic persuaded by Pascal's Wager. Alternatively, one might have evidence for or against p that was too inconclusive to warrant anything more than the judgment that p is possibly true, and yet intend to bring it about that one believe that p. So, there is clearly no sense in which one can only intend to bring it about that one believes p when one holds p to be false. So (A) does not entail B.

Nor does (B) entail (C). I can believe that p is false and intend to bring it about that I acquire the belief that p without it actually being part of the content of my intention that I come to believe a falsehood. That is to say, to put it in a way that parallels a familiar point in epistemology,

intentions are not closed under known logical implication. I can know that x entails y and intend to bring about x without ipso facto intending to bring about y.

The epistemological parallel is worth pursuing. The thesis that knowledge is not closed under known logical implication can be motivated by placing a tracking requirement on knowledge (Nozick 1981). I cannot know that p unless my belief that p tracks the fact that p. That is to say, I cannot know that p unless

(a) were it to be the case that not-p, I wouldn't believe that p

(b) were it to be the case that p I would believe that p.

So, I might know that p entails q and also know that p without knowing that q – because my belief that q, unlike my belief that p, does not satisfy the tracking requirements (a) and (b).

We can place a similar tracking condition upon intentions. Suppose I know that if I go out for a walk on the moor this afternoon it is inevitable that I will end up killing midges (because they are all over the place and I know that I cannot stop myself from crushing them if they land on me, as they are bound to do). Yet I form the intention to go out for a walk and end up killing midges. Have I killed those midges intentionally? It is plausible that I haven't. After all, the killing of the midges was not my aim in going out, nor was it a means to my achieving my aim. My aim was simply to go out for a walk, and even if my going out hadn't resulted in the death of a single midge I still would have gone out.

Here is a case in which it seems that I know that one action entails another and yet intentionally perform the first without intentionally performing the second. The case of self-deception seems exactly parallel. I know that my bringing it about that I come to acquire the belief that p will have the consequence that I come to believe a falsehood. Nonetheless, I can intend to bring it about that I believe that p without intending to bring it about that I believe a falsehood – because in a counterfactual situation in which my bringing it about that I believe that p does not result in my believing a falsehood I would nonetheless still bring it about that I believe that p. In other words, I do not intentionally bring it about that I believe a falsehood because that is not my aim. So (B) does not entail (C).

Although there is no entailment from (A) to (B) to (C) there is an entailment from (C) to (B) to (A). (A) is a minimal requirement on intentionalist theories of self-deception. One can be an intentionalist without accepting (B) or (C), but not without accepting (A). We can see this by a quick comparison with Mele's anti-intentionalist approach to self-deception (Mele 1997, 1998). Mele holds that self-deception occurs when:

- (i) The belief that p which S acquires is false.
- (ii) S treats data seemingly relevant to the truth-value of p in a motivationally biased way.
- (iii) This motivationally biased treatment non-deviantly causes S to acquire the belief that p ,
- (iv) The evidence that S possesses at the time provides greater warrant for $\sim p$ than for p .

Examples of the motivational biases that Mele mentions in spelling out the second condition are: selective attention to evidence that we actually possess; selective means of gathering evidence; negative misinterpretation (failing to count as evidence against p data that we would easily recognise as such were we not motivationally biased); positive misinterpretation (counting as evidence for p data that we would easily recognise as evidence against p were we not motivationally biased), and so forth (Mele 1997). The non-deviant causal chain that leads from motivationally biased treatment of evidence to the acquisition of the belief that p does not proceed via an intention to bring it about that one believes that p .

Let me turn now to what is often taken to be an obvious objection to any intentionalist account of self-deception, but which is actually only applicable to (B) and (C), namely, that such accounts assume that the self-deceiver forms an intention to bring it about that he acquire a belief that he thinks is false. There are two main reasons for holding this to be incoherent. First, one might argue that one cannot intend to bring it about that one acquires a belief that one thinks to be false without simultaneously having contradictory beliefs – the belief that p and the belief that $\sim p$. Yet it is impossible to be in any such state. Second, one might think that the project is guaranteed to be self-defeating. Quite simply, if one knows that the belief is false then how can one get oneself to believe it?

The first of these lines of argument is extremely unconvincing. The key claim that it is impossible simultaneously to possess contradictory beliefs is highly implausible. There is certainly something very puzzling about the idea of an agent simultaneously avowing two contradictory beliefs or avowing the contradictory belief that $p \ \& \ \sim p$. But nothing like this need occur in either (B) or (C), since the two beliefs could be inferentially insulated. Positing inferential insulation is not just an ad hoc manoeuvre to deal with the problem of self-deception (in the way that dividing the self into deceiver and deceived would be), since there are familiar computational reasons for denying that an agent's beliefs are all inferentially integrated (the limitations of memory search strategies etc). In any case, it is a simple logical point that 'S believes p at time t ' and 'S believes q at time t ' do not jointly entail 'S believes $p \ \& \ q$ at time t '. So, an account of self-deception can involve the simultaneous ascription of beliefs that p and that not- p without assuming that those two contradictory beliefs are simultaneously active in any way that would require ascribing the contradictory belief that $p \ \& \ \sim p$.

But there is a sense in which this is peripheral, because it is far from clear that either (B) or (C) do require the ascription of simultaneous contradictory beliefs. I can start from a state in which I believe that $\sim p$ and then intentionally bring it about that I acquire the belief that p without there being a point at which I simultaneously believe that p and believe that $\sim p$. This becomes clearer when one reflects that the best way of bringing it about that one moves from a state in which one believes $\sim p$ to a state in which one believes p is to undermine one's reasons for believing $\sim p$ and hence to weaken one's belief in $\sim p$. It seems plausible that one's confidence in p will be inversely proportional to one's confidence in $\sim p$.

It might be argued that one cannot do something intentionally without doing it knowingly. Hence one cannot intentionally bring it about that one believes p when p is false without knowing that p is false - and so one will have simultaneous contradictory beliefs after all. I shall shortly argue that the premise is false, but let me concede it for the moment. The first thing to notice is that this only threatens intentionalists who espouse (C). Those who espouse (B) are left untouched. Presumably what one knows is the content of one's intention and in (B) the content

of that intention does not include the falsehood of the belief that one is trying to get oneself to believe.

But should we conclude that a (C)-type intentionalist is committed to the ascription of simultaneous contradictory beliefs? Not at all. During the process of intentionally bringing it about that one comes to believe a falsehood one will, on the current assumption, know that one is bringing it about that one will acquire a false belief. But during that time one has presumably not yet acquired the false belief in question. So there is no conflict. And when the process of acquiring the belief has come to an end, so too does the intentional activity and the concomitant knowledge that the belief thus acquired is false. Someone might argue that this will still not allow the self-deceived believer to believe that *p* with impunity – because one cannot believe *p* without believing that *p* is true and one cannot believe that *p* is true if one believes that one has caused oneself to believe it. But this confuses the normative and the descriptive. No doubt one ought not to believe that *p* if one believes that one caused oneself to believe that *p*. But as a simple matter of psychological fact people can reconcile those two beliefs. One might believe, for example, that although one initially set out to cause oneself to believe that *p* the evidence in favour of *p* was so completely overwhelming that one would have come to believe that *p* regardless.

In any case, it seems false that one cannot do something intentionally without doing it knowingly. There is certainly a clear sense in which it is pretty implausible that one might intentionally perform a simple action like pulling a trigger or switching on a light without knowing that that is what one is doing. And it also seems pretty clear that if such a simple action has consequences which one has no means of recognising or predicting (like assassinating the only woman to believe that Arkansas is in Australia or frightening away a tawny owl) then those consequences are not achieved intentionally. But most intentional actions do not fit neatly into one or other of those categories. Suppose I have a long-term desire to advance my career. This long-term desire influences almost everything I do in my professional life, so that it becomes correct to describe many of my actions as carried out with the intention of advancing my career. Does this mean that I carry them out knowing that they are being done with that intention? Surely not.

The intention to bring it about that one acquire a certain belief is closer to the intention to advance one's career than it is to the intention to switch on a light. As Pascal pointed out, acquiring a belief is a long-term process involving much careful focusing of attention, selective evidence gathering, acting as if the belief was true, and so forth. It seems likely that the further on one is in the process, and the more successful one has been in the process of internalising the belief, the more one likely one will be to have lost touch with the original motivation.

It would seem, then, that the standard objections to intentionalist accounts of self-deception are less than convincing. This goes some way towards weakening the case for anti-intentionalism, simply because a considerable part of the appeal of anti-intentionalism comes from the puzzles or paradoxes that are supposed to beset intentionalist approaches. But what about the positive case for intentionalism?

The positive case for intentionalism is based on inference to the best explanation. The intentionalist proposal is that we cannot understand an important class of psychological phenomena without postulating that the subject is intentionally bringing it about that he come to have a certain belief. The situation may be more accurately characterised in terms of one of the three models I have identified - the most common, no doubt, will be (A)-type intentional self-deception. It is not enough for the intentionalist to show that such situations sometimes occur - perhaps by citing extreme examples like the neurological patients who deny that there is anything wrong with them, despite being cortically blind (Anton's syndrome) or partially paralysed (anosognosia for hemiplegia). The intentionalist needs to capture the middle ground by showing that many of the everyday, "common-or-garden" episodes that we would characterise as instances of self-deception need to be explained in intentionalist terms. The task is obviously too large to undertake here, but I will make a start on it by trying to show that the sophisticated mechanisms that Alfred Mele proposes for an anti-intentionalist and deflationary analyses of everyday self-deception look very much as if they can only do the explanatory work required of them when supplemented by an intentionalist explanation.

Let us look again at the four conditions that Mele places upon self-deception:

(i) The belief that p which S acquires is false

- (ii) S treats data seemingly relevant to the truth-value of p in a motivationally biased way.
- (iii) This motivationally biased treatment non-deviantly causes S to acquire the belief that p,
- (iv) The evidence that S possesses at the time provides greater warrant for $\sim p$ than for p.

I would like to focus upon the second condition. Mele requires that S treat data seemingly relevant to determining the truth-value of p in a motivationally biased way. It is presumably a desire that p be the case that results in the motivationally biased treatment. Because we desire that p be the case we engage in selective evidence-gathering, various forms of misinterpretation of evidence, and so forth, eventually resulting in the acquisition of the belief that p. This account is radically incomplete, however. What is the connection between our desire that p be the case and our exercise of motivational bias?

We can get a sense of how Mele would respond from his (1998) discussion of the model of everyday hypothesis testing developed by Trope and Liberman (1996) to explain the divergence. The basic idea is that people have different acceptance/rejection thresholds for hypotheses depending upon the expected subjective cost to the individual of false acceptance or false rejection relative to the resources required for acquiring and processing information. The higher the expected subjective cost of false acceptance the higher the threshold for acceptance - similarly for rejection. Hypotheses which have a high acceptance threshold will be more rigorously tested and evaluated than those which have a low acceptance threshold. Mele proposes that, in many cases of self-deception, the expected subjective cost associated with the acquired false belief is low. So, for example, the complacent husband would be much happier falsely believing that his wife is not having an affair than he would be falsely believing that she was having an affair - because he desires that she not be having an affair. So the acceptance threshold for the hypothesis that she is not having an affair will be low, and it is this low acceptance threshold which explains the self-deceiving acquisition of the belief that she is not having an affair.

We can see how Mele would most likely respond to the question of how the second condition of his account of self-deception comes about. S's desire that p be true results in a motivationally biased treatment of data by affecting the acceptance and rejection thresholds of the hypothesis that

p. It lowers the acceptance threshold and raises the rejection threshold, thus opening the door to biased treatment of the data. This account is ingenious and no doubt provides a good explanation of why different people will draw different conclusions from similar bodies of information. It also provides a good explanation of at least some instances of belief acquisition that one might intuitively classify as instances of self-deception. It is not clear, however, that it can be extended to provide a general account of self-deceiving belief acquisition. There is a fundamental problem that the theory does not seem to address.

Self-deception is paradigmatically selective. Any explanation of a given instance of self-deception will need to explain why motivational bias occurred in that particular situation. But the desire that p should be the case is insufficient to motivate cognitive bias in favour of the belief that p. There are all sorts of situations in which, however strongly we desire it to be the case that p, we are not in any way biased in favour of the belief that p. How are we to distinguish these from situations in which we desire p and are biased in favour of the belief that p? I will call this the selectivity problem.

In response to an earlier presentation of the selectivity problem (Bermudez 1997), and to a related but different problem identified by William Talbott (1995), Mele (1998) gives the following illustration of how his theory might cope with it. He imagines Gordon, a CIA agent who has been accused of treason. Although Gordon's parents and his staff of intelligence agents have access to roughly the same information relative to Gordon's alleged crime, and they all want Gordon to be innocent, they come to different verdicts. Gordon's parents decide that he is innocent, while his colleagues decide that he is guilty. How can this be, given that the intelligence agents and the parents have the same desire and access to similar information? Here is Mele's response.

Here it is important to bear in mind a distinction between the cost of believing that p and the cost of believing falsely that p. It is the latter cost, not the former, that is directly relevant to the determination of the confidence thresholds on Trope and Liberman's view. For Gordon's parents, the cost of believing falsely that their son is innocent may not differ much from the cost of believing that he is innocent (independently of the truth or falsity of the belief). We may suppose that believing that he is innocent has no cost for them. Indeed, the belief is a source of comfort, and believing that Gordon is guilty would be quite painful. Additionally, their believing falsely that he is innocent may pose no subjectively significant threat to the parents. However, with Gordon's staff matters are very different. The cost to them of

believing falsely that Gordon is innocent may be enormous, for they recognise that their lives are in his hands. And this is so even if they very much want it to be true that he is innocent. His parents have a much lower threshold for accepting the hypothesis that Gordon is innocent than for rejecting it, whereas in the light of the relative costs of "false acceptance" and "false rejection" for the staff, one would expect their thresholds to be quite the reverse of this. (Mele 1998 p.361)

In essence, the divergence arises because the CIA agents' desire that Gordon be innocent is trumped by their desire not to be betrayed and their acceptance and rejection thresholds differ accordingly from the threshold of Gordon's parents.

The cost-benefit analysis provides a plausible explanation of what might be going on in the Gordon case, which is particularly interesting since there are ways of describing the situation on which Gordon's parents come out as self-deceived (as indeed there are for describing the CIA agents as being self-deceived). But it is not at all clear, despite what Mele claims, that it provides a response to the selectivity problem. The selectivity problem is not a problem of how two people in similar situations can acquire different beliefs. It arises, rather, from the fact that possessing a desire that *p* be true is not sufficient to generate cognitive bias, even if all other things are equal (which they are, perhaps, for Gordon's parents but not for his subordinates). It is simply not the case that, whenever my motivational set is such as to lower the acceptance threshold of a particular hypothesis, I will end up self-deceivingly accepting that hypothesis.¹ The selectivity problem reappears. There are many hypotheses for which my motivational set dictates a low acceptance and high rejection threshold and for which the evidence available to me is marginal enough to make self-deception possible. But I self-deceivingly come to believe only a small proportion of them. Why those and not the others?.

Intentionalist accounts of self-deception have a clear and straightforward answer to the selectivity problem. The self-deceiving acquisition of a belief that *p* requires more than simply a desire that *p* be the case and a low acceptance threshold/high rejection threshold for the hypothesis that *p*. It requires an intention on the part of the self-deceiver to bring it about that he acquires the belief that *p*. The fact that intentionalist theories can solve the selectivity problem in this way

¹ I leave aside the problem that it seems perfectly possible to deceive oneself into believing a hypothesis

seems at least a prima facie reason for thinking that one cannot entirely abandon an intentionalist approach to self-deception.

Let me end, however, with two qualifications. The first is that the argument from the selectivity problem can only be tentative, because it remains entirely possible that an anti-intentionalist response to the selectivity problem might be developed. It is hard to see what sort of argument could rule this out. Second, even if sound the argument does not compel recognition of the existence of anything stronger than what I have called (A)-type self-deception. I have suggested that (B)- and (C)-type self-deception are not conceptually incoherent. But it remains to be seen whether there are situations for which inference to the best explanation demands an analysis in terms of (B)- or (C)-type self-deception.

Bibliography

- Bermúdez, J. L. 1998. 'Defending intentionalist accounts of self-deception', Behavioral and Brain Sciences 20, 107-108.
- Mele, A. 1997. 'Real self-deception', Behavioral and Brain Sciences 20, 91-102.
- Mele, A. 1998. 'Motivated belief and agency', Philosophical Psychology 11, 353 – 369.
- Nozick, R. 1981. Philosophical Explanations. Cambridge MA. Harvard University Press.
- Talbott, W. 1995. 'Intentional self-deception in a single, coherent self', Philosophy and Phenomenological Research 55, 27-74.
- Trope, Y. and Liberman, A. 1996. 'Social hypothesis testing: cognitive and motivational mechanisms' in E. Higgins and A. Kruglanski (eds.), Social Psychology: Handbook of Basic Principles. New York. Guilford Press.