Abstracta 5 : 2, pp. 126 – 138, 2009

PROSPECTS FOR AN INTENTIONALIST THEORY OF SELF-DECEPTION

Kevin Lynch

Abstract
A distinction can be made between those who think that self-deception is frequently intentional and those who don't. I argue that the idea that self-deception has to be intentional can be partly traced to a particular invalid method for analyzing reflexive expressions of the form '*V*ing oneself' (where *V* stands for a verb). However, I take the question of whether intentional self-deception is possible to be intrinsically interesting, and investigate the prospects for such an alleged possibility. Various potential strategies of intentional self-deception are examined in relation to Alfred Mele's suggestion that doing something intentionally implies doing it knowingly. It is suggested that the best prospects for an intentionalist theory of self-deception lie with a strategy involving the control of attention.

1. Two approaches to the analysis of self-deception

The self-deception debate is riven by a theoretical divide between so-called 'traditionalists' and 'deflationists' (though many philosophers take up mixed positions between them). These theoretical differences can be traced in large part to two different approaches to how the concept of self-deception should be properly analyzed, usefully distinguished by Alfred Mele. One, he calls the *lexical* approach, the other, the empirical or *example-based* approach. On the lexical approach; '[w]e might start by asking what "deception" (or "deceive") means, and then ask what "self-deception" must mean if it is to be a species of deception'.[1] Alternatively, on the example-based approach; '[o]ne starts by gathering and constructing cases that would generally be described as self-deception, and then attempts to develop an analysis of self-deception on the basis of a consideration of this material. The meaning of "self-deception" is determined by the cases, which are therefore the most fundamental data'.[2] Note that these two approaches are somewhat idealized, and it can be difficult to find a philosopher who explicitly and exclusively adopts one.

---

[1] Mele (1987: 13).
[2] Mele (1987: 13-14). After making this two-fold distinction in a 1987 paper, he later introduced another category; the 'theory-guided approach'. As he defines it, on this approach 'the search for a definition is guided by commonsense theory about the etiology and nature of self-deception (1997: 92). However, Mele

Both approaches yield quite different answers to the question of what self-deception must be. Mele himself practices the example-based approach. He thinks that the meaning of a term of folk-psychology is a function of how ordinary folk use it.[3] Accordingly, he starts by considering typical cases in which we would pre-theoretically refer to someone as having deceived him/herself: the terminally ill patient in denial, the husband who won't believe that her wife is having an affair when it should be obvious, the mother who won't believe that her son is taking drugs, etc. What goes on in such cases, Mele argues quite persuasively, is that judgment becomes distorted and biased by desire and emotion in ways that the subject *didn't intend*.

On the lexical approach, the characterization is different. Basically, this approach starts by establishing a definition of deception from the interpersonal case, and uses it to deduce the meaning of 'self-deception'. Therefore, the meaning of 'self-deception' can be established, on this view, *independently* of looking at or taking into account how ordinary people actually use the expression 'self-deception', and thus independently of any study of the 'garden-variety cases' that Mele speaks of (that the lexical approach may alienate the philosopher from the actual use of this word will be made clearer shortly).

I think that we can understand this approach as being guided, whether explicitly or implicitly, by the following formula. We can call it the 'lexical formula':

> What it means for someone to deceive himself is for him to do the same thing to himself that he does to another when he deceives another.

To many an ear this formula may sound intuitively compelling. And it seems to give us the right result in many instances that spring to mind. For example, if Jones shoots Smith, Jones points a loaded gun at Smith and pulls the trigger. And what else, in that case, could it be for Jones to shoot himself, if not to do that same thing, but to himself, namely; point a loaded gun at himself and pull the trigger?

---

doesn't say anything about what these 'commonsense theories' are, and it is unclear what philosophical accounts he has in mind as exemplifying this approach.

[3] Mele (1998: 39).

On the lexical approach, before we can employ this formula to see what it is to deceive oneself, we must first establish what it is to deceive another. The following definition is usually taken to capture what this is:

> When A deceives B, A intentionally/deliberately causes B to believe something that A knows/suspects is false.

It is true that some clever counterexamples have been advanced against this definition, e.g. Barnes.[4] However, almost all the philosophers, including Mele,[5] agree that this definition captures the conceptually central or stereotypical cases of interpersonal-deception. So the traditionalist does have scope to argue that the counterexamples are conceptually peripheral or limiting cases of deception, and for that reason she can adhere to this definition using the qualifier 'paradigmatically'. Then, feeding this definition into our formula, we can deduce that paradigmatically:

> When A deceives himself, A intentionally/deliberately causes himself to believe something he knows/suspects is false.

Now if we assume the validity of the lexical formula, and of the definition of deception derived from the interpersonal paradigms, then it follows logically that self-deception must be as this definition says. On this picture of self-deception, A, after encountering evidence that makes him realize that some unwelcome proposition *p* is true, deliberately causes himself to believe the contrary, welcome proposition not-*p* (perhaps to avoid the anxiety of knowing that *p*). Some traditionalists argue that we also get the result that A ends up in a condition where he believes that *p* and believes that not-*p* simultaneously. However, it's not obvious why the lexical derivation as it stands would necessarily imply this. Traditionalists here typically advert to the fact that as deceiver, A must believe that *p*, and as deceived he must believe that not-*p*, but this only begs the question of why the person must satisfy both

[4] Barnes (1997: 8-11).
[5] Mele (1997: 92).

these roles simultaneously, rather than consecutively.[6] It may be that the traditionalist needs some additional assumptions added to this 'lexical' analysis to support this controversial aspect of his/her account.

Nevertheless, on the whole it seems that if we grant the lexical approach, then there is a pretty strong case to be made for the thesis that self-deception must be as this definition states, which by-and-large amounts to the classic traditionalist account. However, it would not follow from this that this phenomenon actually exists. On the lexical approach, the philosopher asks *hypothetically*; what conditions would *have to* be met for there to be a case of self-deception. For that reason, there is room for the question of whether self-deception ever obtains at all or whether it is even possible, and as Mele points out, many who take the lexical approach end up being skeptics about self-deception.[7] Note that this question is ruled out from the outset by the Mele-type approach, since Mele's methodology takes the legitimacy of people's customary use of 'self-deception' for granted and just asks what goes on in the cases so referred to.

2. Problems with the lexical approach

As I've said, if we grant the lexical approach we get a strong case for the traditionalist account. However, there appear to be difficulties with this methodology. The problem lies with the lexical formula. We should expect that we could turn this formula into a general formula for deriving the meaning of any reflexive construction grammatically analogous to 'deceiving yourself'. Accordingly, we can state the lexical formula in general terms as follows:

> What it means for one to *V* oneself is for one to do the same thing to oneself that one does to another when one *V*s another.

[6] Though one might point out that in the interpersonal cases A knows that *p* when B acquires the belief that not-*p*, though typical, this is not a necessary element of interpersonal deception. For instance, A could send a deceptive letter to B and die while it's in transit (see Siegler 1963: 35).
[7] Mele (1997: 92).

…where *V* stands for some verb. But as T.S Champlin has shown,[8] such an approach can be parodied for a number of such reflexive constructions.
Consider how the lexicalist reasoning would work for 'teaching yourself':

> - Teaching yourself is doing the same thing to yourself that you do to another when you teach another.
> - If A teaches B about *x*, A knows about *x* and imparts this knowledge to B.
> - Therefore, if A teaches himself about *x*, A knows about *x* and imparts this knowledge to himself.

Note that our use of the lexical formula has landed us with a 'paradox of self-teaching' analogous to the notorious 'paradox of self-deception', for it seems as though the one who teaches himself must, as teacher, know about *x* and at the same time, as student, be ignorant of *x*.

As regards analyzing the notion of being self-taught, the different consequences that would ensue from adopting the lexical approach, as opposed to the example-based approach, are clear and striking. Adopting the former, we get the outlandish result that teaching yourself is imparting your own knowledge to yourself. But this answer clearly doesn't tally with the actual use of the expression. This use is simply not constrained by the strictures of any such formula. The cases we refer to with that phrase are not cases in which people impart knowledge to themselves (if that is intelligible). They are cases in which people *lack* the relevant knowledge, but acquire it by solitary study, trial and error, and so on, without the benefit of a teacher. Adopting the example-based approach would give an answer like that to the question of what it is to be self-taught. Adopting the lexical approach, we would probably end up being skeptics about the possibility of teaching oneself, or we might construct elaborate metaphysical theories to explain how it is possible to impart knowledge to yourself (mental 'divisions' and 'sub-systems', etc.), all of which would be a surreal reflection of the self-deception debate.

[8] Champlin (1977: 284-285. 1988: 24-25).

It also seems clear that the lexical approach gives us the *wrong* answer for this case. For nobody wants to say that in our ordinary use of that expression we are *mis*using it, because we are referring to cases that are not cases of imparting knowledge to oneself. Therefore, the application of the lexical formula to the case of self-teaching constitutes a *reductio ad absurdum* of the lexical approach, conceived of as a general approach to the analysis of constructions of the form 'self-*V*'.

Nevertheless, despite the fact that it seems to be the offspring of an invalid analytic method, the traditionalist definition of self-deception has created an interesting philosophical research-program into the bounds of the possible, since some philosophers claim that if self-deception were as this definition states, then deceiving yourself would turn out to be impossible to do (at least under normal circumstances), while others claim that not only is this phenomenon possible, but it's a common occurrence. Therefore–despite the shortcomings of the lexical approach–I want to *suppose* the lexicalist definition of self-deception, *just for the sake of argument*, to see whether it gets us a possible phenomenon. We can do this so long as we bear in mind that the investigation may have little relevance to our understanding ordinary self-deception, just as the lexical analysis of self-teaching has little relevance for our understanding of ordinary self-teaching.[9]

3. The obstacle to intentionally deceiving oneself

Skepticism may arise over this notion of self-deception because of the requirement for it to be intentional or deliberate (terms which philosophers often use interchangeably here). The worry is well-put by Mele:

> It is often held that doing something intentionally entails doing it knowingly. If that is so, and if *deceiving* is by definition an intentional activity, then one who deceives oneself does so *knowingly*. But knowingly deceiving oneself into believing that *p* would require knowing that what one is getting oneself to believe is false. How can that knowledge fail to undermine the very project of deceiving oneself?[10]

[9] Traditionalists do, however, have other resources for arguing that ordinary self-deception is often intentional, including arguments to the best explanation.
[10] Mele (1997: 92).

Mele's presupposition is that if one *V*ed intentionally, one necessarily must have *known* or *been aware* that one was doing *that*, namely; *V*ing.[11] This seems to suggest that 'unconsciously yet intentionally *V*ing' is contradictory. Let's call this the *knowledge condition* (though we could equally well call it the 'awareness condition'). Philosophers have frequently thought that there's a conceptual connection between the notions of intentional action and knowledge/awareness of what you're doing.[12] Some lexicographers apparently assume so too. For instance, the *American Heritage Dictionary*, 4th edition, defines 'deliberate' as 'done with or marked by full consciousness of the nature and effects; intentional'. I assume the relevant considerations that might support such claims would be as follows. Take any occasion when you do something without knowing or being aware that you are doing it. Say, for instance, that you are making soup, and you unwittingly add some sugar (you think that it's salt). Or imagine that you travel to an exotic country with very different customs and you meet one of the locals. You do something that would not attract any notice in your country but which is taken to be highly insulting in this culture, causing great offence to the local, though you weren't aware that this action is considered offensive. Now it seems clear that in these cases we added the sugar or offended the local neither intentionally nor deliberately, and in justifying this, we naturally advert to the fact that we weren't aware that we were doing that. And it's difficult to see how the accusation that we did these things intentionally/deliberately could stand given this lack of awareness. A philosopher could then argue that by parity with such cases, if someone deceived herself intentionally, she must have known that she was doing that.

Though the connection between intention and knowledge/awareness seems to be intimate, there is no opportunity here to investigate whether it allows for exceptions to the knowledge condition. I am just going to grant this condition from here on in. Let's just say that the onus seems to be on the philosopher who assumes that the intentional self-deceiver

[11] Actually, whether Mele is really committed to this is unclear, since he states elsewhere that 'hidden intentions' are possible (1997: 100). Nevertheless, Mele does think that there is something paradoxical about the idea of intentionally deceiving yourself, and it's hard to see what else could be generating this paradox if not this presupposition.

[12] See Anscombe (1957: 11 & 87), Miller (1980: 334), Gustafson (1975: 89), Bratman (1984: 387), Hampshire (1970: 145), Donnellan (1963: 406), Moran (2001: 125) and Hamlyn (1971: 46).

is not aware of what she's doing to explain how this is possible.[13] So how exactly does the knowledge condition constitute an obstacle to intentional self-deception?

Let's consider a number of strategies–strategies that could be used to deceive another–and see if they could be used to deceive oneself. Strategies of self-deception can for our purposes be defined as methods for bringing about *deviant* doxastic changes in ourselves, where 'deviant' is supposed to exclude the legitimate ways in which we may bring about belief-changes in ourselves (e.g. acquiring crucial new evidence, noting a mistake in one's reasoning, etc.). Typically, philosophers are interested in self-reliant strategies that may be available for use in ordinary circumstances. So strategies such as hiring a hypnotist or a brain-scientist to accomplish the goal, though possibly effective, are usually not considered interesting (perhaps because they are not contenders for explaining what's happening in ordinary cases of self-deception). Examples of such strategies often mentioned in the literature include the following:

1) Lying to yourself
2) Manipulating and distorting the evidence
3) Rationalizing
4) Selective gathering of evidence

Let me consider (1) and (2) first. As I understand it, Mele's argument would take the following form: 'If you lie to yourself intentionally, then (assuming the knowledge condition), you know that *that* is what you are doing, namely, uttering a *lie* (i.e. an untruth). For you to know that would presumably render you immune to it.' Likewise, 'If you distorted the evidence intentionally, then you are aware that you have done *that*, namely, distorted the evidence. Therefore, you couldn't be taken in by it.'

Strategy (3) is trickier. Whether this type of argument works will entirely depend, I believe, on whether 'rationalize' is pejorative. On Kent Bach's definition, rationalization is 'any case of a person's explaining away what he would normally regard as adequate

[13] Too often, unconsciously intentional action is attributed to the self-deceiver more in an *ad hoc* manner to save the theory of intentional self-deception, than on the basis of any independent considerations (e.g. Steffen 1986, p.47).

evidence for a certain proposition'.[14] Let's assume here that 'explaining away' is pejorative, so that it means something like 'giving bogus arguments' to undermine evidence. Assuming this, if Jones rationalizes intentionally, and if rationalizing means adducing bogus arguments, then Jones intentionally adduces bogus arguments. And if doing that intentionally implies doing it knowing that *that* is what one is doing, then he *understands* those arguments to be *bogus*. Therefore, he couldn't be taken in by them. Though someone may intentionally rationalize to fool another, the idea of intentionally rationalizing to fool *yourself*–given our assumptions–is incoherent. One would expect rationalizers not to understand themselves to be rationalizing, and so not to be deliberately rationalizing.

Similar arguments apply to (4). To accuse someone of *selectively* gathering evidence is to accuse someone of having done something unjust. To know that one gathered evidence selectively is to know that one did not show justice to both sides of the argument, neglecting one of the sides, and so is to know the dubious value of one's efforts. So granting the knowledge condition, self-deceivers cannot intentionally/deliberately gather evidence selectively.

So we see the form of Mele's argument. A pejorative term is used to denote some epistemically untoward activity. The fact that the strategy is executed intentionally, granting the knowledge condition, implies that the agent knows that his activity is untoward. This knowledge would then preclude the agent being fooled by the product of this activity.[15]

4. The attentional strategy

However, there are theories of self-deception which mention certain kind of actions that may be intentionally executed, and which may be thought to have potential as an effective means for deceiving oneself, *even granting* the knowledge condition. Robert Lockie

[14] Bach (1981: 358).

[15] On some accounts of the strategy of intentional self-deception, the self-deceiver might intentionally do something designed to deceive his/her future self. Here it would be the knowledge of what one *did*, rather than of what one *is doing*, that would be the obstacle to success.

usefully categorizes this type of account under the label 'attentional accounts'.[16] We can formulate the idea as follows:

> A person who believes or at least suspects that *p*, but who wishes to believe that not-*p*, by turning his attention away from the unwelcome considerations supportive of *p*, and by attending to welcome considerations supportive of the contrary not-*p*, may end up losing the (conscious) belief that *p* and acquiring the belief that not-*p*.

So self-deception, on this account, involves what psychologists call *thought-suppression*, i.e. the act of ridding thoughts from the mind–as well as selective focusing of attention. The idea may be illustrated with the following. Harry is a goal-keeper who has lately been letting in some easy shots. He's beginning to think that he's not a very good goal-keeper and this distresses him. The attentional theory states that by avoiding the thoughts of his poor performances and by concentrating on the few memories of when he performed well, Harry may come to believe that he's a good goal-keeper, even though the totality of the evidence that he was acquainted with suggests that he is not.

Now why would this strategy be thought to overcome the knowledge condition? The answer, I believe, would be assumed to lie with the thought-suppression element of the strategy. Thought-suppression may be conceived of as a *knowledge-subverting* strategy. For example, Harry shifting his attention off the memories of his bad performances is designed to undermine the knowledge that gets in the way of his having the welcome belief. How this happens, in the opinion of Van Leeuwen[17] and Whisner,[18] is that after perhaps repeated attempts at thought-avoidance Harry supposedly *forget*s about those performances.

The knowledge condition might still be considered as a stumbling block, however. If the thought-suppresser intentionally shifts her attention off a thought, doesn't this mean that she will be intentionally not thinking about that thought? But granting the knowledge condition, this would imply that she's aware of what she's doing, i.e. not thinking about it.

[16] Lockie (2003: 131).
[17] Van Leeuwen (2008: 202).
[18] Wisner (1998: 196).

And paradoxically, this seems to imply that the thought is in mind, so that it has not really been forgotten after all.[19]

But there is little reason to entertain such worries. Here it may be useful to look at what thought-suppression typically involves. Daniel Wegner–one of the most prominent researchers on this issue–says that when people try to take their mind off something, they typically do so by putting it onto something else.[20] They turn their mind or attention to a 'distracter' which may absorb their attention. Now although this distraction seeking can be done intentionally, this doesn't imply that the agent is intentionally not thinking about the unwelcome evidence when she succeeds in distracting herself from thinking about it. We can look on the action of suppressing a thought as analogous to the action of going to sleep. In both cases, we intentionally do things to *facilitate* certain results, i.e. certain thoughts being lost from consciousness, or our losing consciousness altogether. We may intentionally facilitate these results in the latter case by lying down in a dark, quiet room, and in the former, by shifting our attention onto something else. But the fact that we intentionally tried to suppress a thought doesn't imply that when the thought is forgotten, we know that it is, or that we are intentionally not thinking about it, any more than our intentionally going to sleep implies that when we finally fall asleep, we know (or are aware) that we are asleep, or that we are intentionally sleeping.[21] These results are things that we intended without being things we are doing intentionally.

The 'attentional account' of self-deception, though popular, remains rather under-elaborated in the literature, and it remains to be seen whether it could represent a realistic strategy of self-deception. However, I believe it offers perhaps the best prospects for an intentionalist theory. The reason is that with the other strategies usually mentioned, nothing is done to subvert the knowledge associated with the deliberateness of the attempt, knowledge that would render the attempt futile. Suppressive strategies, however, are supposedly knowledge-subverting: they may be aimed at undermining knowledge of the considerations that support the unwelcome belief, perhaps by undermining memory of

[19] For similar worries, see Pugmire (1969: 346) and Reilly (1976: 393).
[20] Wegner (1994: 12 & 60).
[21] Suicide is another example. One commits suicide intentionally, but can't know that one has succeeded when one has.

those considerations. Whether we have those abilities for mental manipulation, however, is debatable,[22] and it may ultimately be an issue for psychologists to have the last say on.[23]

Kevin Lynch

Warwick University

Kevinlynch405@eircom.net

References

Anscombe, G.E.M. (1966) *Intention*, Oxford: Basil Blackwell.

Bach, K. (1981) 'An Analysis of Self-Deception', *Philosophy and Phenomenological Research*, 41, 351-370.

Barnes, A. (1997) *Seeing Through Self-Deception*, Cambridge: Cambridge University Press.

Bratman, M. (1984) 'Two Faces of Intention', *Philosophical Review*, 93, 375-405.

Champlin, T.S. (1988) *Reflexive Paradoxes*, London: New York: Routledge.

Champlin, T.S. (1977) 'Self-Deception: A Reflexive Dilemma', *Philosophy*, 52, 281-299.

Donnellan, K.S. (1963) 'Knowing What I am Doing', *Journal of Philosophy*, 60(14), 401-409.

Gustafson, D. (1975) 'The Range of Intentions', *Inquiry*, 18, 83-95.

Hamlyn, D.W. (1971) 'Self-Deception', *Proceedings of the Aristotelian Society* (Supplementary volume 35), 45-60.

[22] Wegner's empirical studies of people's attempts to suppress unwanted thoughts (1994) have highlighted how difficult people find it to rid themselves of an unwanted thought or make themselves forget an unwanted memory.

[23] Thanks to Johannes Roessler, and to the audiences of the Second European Graduate School, Bochum Germany, and the Warwick philosophy department WIP seminar group, for useful comments on this paper. This paper was supported by a Warwick Postgraduate Research Scholarship.

Hampshire, S. (1970/1959) *Thought and Action*, London: Chatto and Windus.

Lockie, R. (2003) 'Depth-Psychology and Self-Deception', *Philosophical Psychology*, 16, 127-148.

Mele, A.R. (1998) 'Two Paradoxes of Self-Deception', IN Jean-Pierre Dupuy (ed.) *Self-Deception and the Paradoxes of Rationality*, Stanford: CSLI Publishing, 37-58.

Mele, A.R. (1997) 'Real Self-Deception', *Behavioral and Brain Sciences*, 20, 91-102.

Mele, A.R. (1987) 'Recent Work on Self-Deception', *American Philosophical Quarterly*, 24, 1-17.

Moran, R. (2001) *Authority and Estrangement*, Princeton; Oxford: Princeton University Press.

Miller, A.R. (1980) 'Wanting, Intending, and Knowing What One is Doing', *Philosophy and Phenomenological Research*, 40, 334-343.

Pugmire, D.R. (1969) '"Strong" Self-Deception', *Inquiry*, 17, 339-361.

Reilly, R. (1976) 'Self-Deception: Resolving the Epistemic Paradox', *The Personalist*, 57, 391-394.

Siegler, F.A. (1963) 'Self-Deception', *Australasian Journal of Philosophy*, 41, 29-43.

Steffen, L.H. (1986) *Self-Deception and the Common Life*, New York etc.: Peter Lang.

Van Leeuwen, D.S.N. (2008) 'Finite Rational Self-Deceivers', *Philosophical Studies*, 139, 191-208.

Wegner, D.M. (1994) *White Bears and Other Unwanted Thoughts*, New York; London: The Guilford Press.

Whisner, W. (1998) 'A Further Explanation and Defense of the New Model of Self-Deception: A Reply to Martin', *Philosophia*, 26, 195-206.