Pre-print of paper published in *Erkenntnis* 78(6), pp.1337-1345.

Self-Deception and Stubborn Belief

Kevin Lynch

Abstract

Stubborn belief, like self-deception, is a species of motivated irrationality. The nature of stubborn belief, however, has not been investigated by philosophers, and it is something that poses a challenge to some prominent accounts of self-deception. In this paper, I argue that the case of stubborn belief constitutes a counterexample to Alfred Mele's proposed set of sufficient conditions for self-deception, and I attempt to distinguish between the two. The recognition of this phenomenon should force an amendment in this account, and should also make a Mele-style deflationist think more carefully about the kinds of motivational factors operating in self-deception.

1. Mele's Motivated Bias Account of Self-Deception

Alfred Mele has proposed a set of conditions which he claims are sufficient for self-deception and which are supposed to capture the paradigm cases. Critics, however, have objected to this account a number of times over the years with putative counterexamples, describing phenomena which we would regard as distinct from self-deception but which seem to satisfy these same conditions.

I believe that previous arguments of this form have not been successful, and that an account of this sort proves to be resilient in its ability to distinguish self- deception from kindred phenomena. For instance, some have objected to accounts of self-deception in terms of motivationally biased belief as failing to distinguish wishful thinking from self-deception (e.g., Gardiner 1969–1970: 242). Subsequent discussions of that distinction, however, especially that of Szabados (1973, 1974), indicate how an account like Mele's would give us the resources for

distinguishing the two.[1] Another example of this kind of argument is one by Richard Holton, who discusses a case of prejudicial belief and says that it meets Mele's conditions while not counting as self-deception. I will discuss this argument briefly below to show that Holton doesn't succeed in producing a counterexample.

Nevertheless, I believe that there is one phenomenon which hasn't been considered yet which does satisfy his conditions but which seems to be distinct from self-deception, thus making it a genuine counterexample. Recognition of this phenomenon would force an amendment in this account, and should make a Mele-style deflationist think more carefully about the kinds of motivational states causing the bias in self-deception. This phenomenon is stubborn belief, and in this article I will give an account of how a deflationist about self-deception should distinguish between a stubborn believer and a self-deceiver.

This issue is worth pursuing because Mele's set of conditions, in my view, comes close to answering the question of what (paradigmatic) self-deception is,[2] or at least it is one of the most prominent contenders in circulation, and it is thus worth putting to the test and, if appropriate, refining in light of objections. One important way of getting a clearer view of the nature of self-deception, or of anything else for that matter, is by comparing it with kindred phenomena and trying to discern their distinguishing features. Stubborn belief is one such phenomenon, which to my knowledge has not been compared with self-deception yet, and doing this may help put self-deception's contours into greater relief. Stubborn belief, moreover, is something worth studying in its own right for anyone interested in forms of human irrationality and the ways in which affect can unduly influence cognition.

[1] Szabados' idea is that wishful thinkers believe something unwarrantedly because they want it to be true, without having significant evidence either for or against it, often jumping to the welcome conclusion on the basis of a modicum of welcome evidence, while self-deceivers hold their belief *in the face of contrary, unwelcome evidence*. Thus self-deception involves *taking a defensive reaction towards and resisting evidence* (e.g., by reinterpreting it and explaining it away), while such measures aren't necessary in wishful thinking. Condition 4 (see below) of Mele's account, in that case, is critical for excluding cases of wishful thinking. This is why I think that Mele is wrong to say that condition 4 is expendable (2001: 51-52). In fact, it's required if his account is not to conflate these two phenomena, a point he himself hints at elsewhere (1997: 100).

[2] For further efforts to defend this general view of self-deception, see (Lynch 2012).

First let's look at Mele's sufficient conditions for self-deception, though here I word them slightly differently.[3] For Mele, a subject is self-deceived in believing that *p* when:

1. The belief that *p* which *S* has is false.
2. *S* treats data relevant, or at least seemingly relevant, to the truth value of *p* in a motivationally biased way.
3. This biased treatment is a non-deviant cause of *S*'s having the belief that *p*.
4. The body of data possessed by *S* at the time provides greater warrant for not-*p* than for *p* (Mele 2001: 120).

I hope that these points are not in need of much comment, though I will say that by 'motivationally biased' in 2, Mele just means that *S* treats data in a biased way because of a desire (or aversion) which he has.

2. Past Criticism of the Account

As I've said, Mele's account, in my view, holds up well against previous attempts at refutation by counterexample. One such recent attempt was by Holton (2001). Holton argues that Mele's conditions fail to distinguish self-deceptive belief from distinct cases of prejudicial belief, illustrating his point with the following case:

> Jean-Marie is a racist. He thinks that blacks and Arabs are not as good as whites: not as clever, or as imaginative, or as brave, or as trustworthy, or whatever. Take just about any property that Jean-Marie might regard as a virtue, and he will think that whites have more of it. Let us assume that his beliefs here are, by and large, false. But he holds them sincerely. And were we to challenge them, he would provide evidence: reams of it, taken

[3] Mele holds that one can self-deceptively *acquire* the belief that *p*, or self-deceptively *retain* the belief that *p*, and in his exact formulation of these conditions, he presents them as applying to self-deceptively acquiring the belief that *p* by, for instance, stating (1) as 'The belief that *p* which *S* acquires is false'. I have substituted 'has' for 'acquire', which gives the conditions increased generality by being ambiguous on whether the belief was self-deceptively acquired or retained.

> from the magazines and newspapers of the kinds of organisation to which he belongs. He is aware of the opposing view; indeed he has reams of [evidence supporting the other side] too, collected to document the conspiracy which he thinks pervades the liberal establishment that controls the mainstream press and publishing houses (Holton 2001: 59).

Holton says that ‘‘Jean-Marie meets all four of Mele’s conditions. He is bigoted and prejudiced. Yet he is not obviously self-deceived’’ (2001: 59).

The problem here is that it’s not made evident from Holton’s description of the case that it satisfies all of Mele’s conditions, and if that were made clear, it is not so evident that we would be as reluctant to think of it as a case of self-deception as Holton assumes (for saying that a belief is bigoted or prejudiced does *not* preclude that it is *also* self-deceptive). Included among Mele’s conditions is the point that the subject treats the data in a *motivationally biased way* (i.e., he treats it in a biased way because of a desire), which makes him have the belief. But although there is a suggestion that Jean-Marie has engaged in biased reasoning and evidence searching about the issue, there is no indication that this behaviour was driven by Jean-Marie’s hopes and fears. For instance, if it were made explicit that Jean-Marie strongly *desires* that blacks and Arabs are inferior to whites, and that this was what was motivating his biased behaviour which led to him believing that they are inferior, then a diagnosis of self-deception would begin to look much more plausible.

Generally speaking, an explanation in terms of motivation is not a conceptually central feature of prejudicial or racist beliefs. The existence of a prejudicial belief may not even be the outcome of any biased reasoning as such, and may often be explained with reference to various non-motivational influences, such as cultural influences. Such beliefs may get ingrained in one through the influence of associates, of education, customs, or through features of the social structure. They may also arise from indoctrination. In such cases, a desire that the belief is true, or any other desire on the believer’s part, need not feature in the explanation of how the belief was acquired at all, and furthermore, neither may the belief have resulted from biased reasoning (the belief may simply have been adopted from these influences and authorities, without much reasoning having taken place). On the other hand, desires and related affective states *do* play a central role in the explanation of *self-deceptive* beliefs. So if it were true that the racist’s belief

that blacks are inferior was caused or sustained by his desiring that they are (perhaps because he has a personal stake in a social system legitimized through that assumption, as did the white beneficiaries of Apartheid or slavery), where this desire drove him to reason in a biased way about the issue, then self-deception would be a better diagnosis of the situation, and indeed, it is quite plausible that in many instances, racist beliefs *are* sustained by self-deception.

However, there is another phenomenon which satisfies Mele's conditions which, I think, we would want to distinguish from self-deception. In that case, satisfying Mele's conditions will fail to guarantee that any such case is one of self-deception. This phenomenon is stubborn belief.

3. Stubborn Belief

I offer the following as an initial working-definition of a person who is being stubborn in believing something. Such a person is one who, after having settled on some opinion, refuses to reconsider it and sticks to that opinion, when other people are with good reason encouraging him to reconsider, or when he is confronted with evidence which would warrant such a reconsideration and perhaps rejection of that belief. The stubborn believer won't be talked out of his view, and is unyielding or resistant to reasonable persuasion. He clings to the belief, as we say, stubbornly. People find that when they try to criticize or persuade him, he just 'won't listen' and is dismissive of their objections and advice. He ends up holding on to his belief irrationally, after failing to take on board considerations that would warrant believing otherwise.[4]

Note that often we speak of people as stubbornly holding to *courses of action*, rather than beliefs (e.g., a stubborn general who refuses to reconsider his strategy amid signs of impending disaster and the doubts of advisors). However, I suspect that in many of these cases, the subject is stubbornly sticking to the course of action because he is stubbornly clinging to the *belief in the rightness* of the course of action, so these cases will involve stubborn beliefs too.[5]

[4] Note that I don't want to suggest here that it's always unreasonable not to reconsider your beliefs when others are challenging them or encouraging one to reconsider. In the cases being considered here, we are assuming that the challenger's position is warranted and persuasive, and would enjoin the subject to change his mind. Thus, the phrase 'with good reason' is important in the above formulation, which is supposed to indicate that the belief which the subject clings to is not epistemically warranted by his evidence in these cases.

[5] Stubbornness may not always concern belief, however. Consider an old woman whose house is in the projected path of a new motorway, who refuses to relocate despite the fact that all her neighbors have moved and the very

Why do some people stubbornly cling to certain beliefs in the face of new, contrary evidence or reasonable challenge? It would seem to be because of a desire or aversion which they have, but what kind of desire or aversion? Consider the following two cases.

Burke is a proud man who has trouble admitting when he's wrong. This happens especially on occasions where he feels that he has staked his reputation on something, for instance, as a capable manager in his job, or as the reliable man of the house at home. If he makes mistakes in these circumstances, he will do his best to find someone or something else to blame other than himself. He is also a competitive character, and when he gets into debates or disputes with his friends, he hates being outsmarted in an argument or shown up as being wrong, which injures his pride. Thus when criticised by certain people he will often make exorbitant efforts to defend his position, ignoring the virtues of his opponent's case, intent more on winning the argument than establishing the facts.

Or consider Murphy. Murphy is set in his ways. He has his own rather conventional views about how the world works, and he is comfortable with the certainty they give him. He doesn't like it when his long-held views are challenged, because this threatens to take his comfortable certainty away. He is also intellectually quite lazy, and dislikes his worldview being challenged because of the effort involved in having to reconsider the matter all over again if refuted. Consequently, to avoid this effort and to maintain the comfort and security of certainty, he often doesn't give others a proper hearing when they challenge certain opinions of his and finds ways to dismiss their criticisms.

I am sure we all have at least a little bit of Burke and Murphy in us, but it seems to me that these would be examples of people of stubborn disposition, and of the psychological profiles which can sustain such stubbornness. Moreover, if we consider any one occasion where Burke or Murphy are being stubborn in retaining some belief, *p*, it seems that such a case may meet all of Mele's conditions. Firstly, the belief may be false. Moreover, it may be maintained due to a *motivationally biased treatment of the data*. This means that the person concerned may be dealing with the data (i.e., the critical advice or considerations being put to him) in a biased way, due to a desire or aversion which he has. Burke, for instance, hates being shown to be wrong, and this may motivate him to fail to appropriately consider or take on board the considerations of a

---

generous offers of compensation. We may describe her as stubborn, though this may not imply that she is being stubborn with respect to one of her beliefs.

critic, and to find ways to dismiss her points. Or Murphy may hate the disruption involved in losing one of his beliefs, and this may cause him to seek ways to dismiss contrary considerations when it is challenged. Furthermore, these critical considerations may in fact warrant the belief that not-*p* (satisfying condition 4). Nevertheless, it seems that we treat such cases as a distinct form of irrationality to self-deception. We would say that this person is being stubborn, and not self-deceived, in believing that *p*.

Note that we often speak of stubbornness as a *character trait*, and I have tried to reflect this in the cases I have described. In the two cases mentioned, the subjects were described as having aversions which incline them to stubbornly cling to different beliefs on different occasions. Their possession of a certain trait is thus explained by those aversions. However, it may be that a person can be stubborn on a particular occasion without being a generally stubborn person.

Suppose that Flanagan meets Burke at a party, two men with an interest in history. The topic turns to Napoleon, and Flanagan makes plain his view that Napoleon was a warmonger and a ruthless tyrant who wanted to dominate Europe. Burke disagrees, and argues that Napoleon had war forced upon him and that he didn't have the character of a tyrant. Assume that Burke's arguments are sound. Now Flanagan is generally a fair-minded man who is more concerned with truth than with winning an argument, but on this occasion he acts out of character. Something about Burke, his smug, obnoxious, or cocksure manner, riles him so much that he is overcome with an intense desire not to concede to him. So he meets Burke's points with some dismissive comments and specious replies, and privately in his own mind, he deploys *ad hominem* arguments ('if it came from that fool, it can't be right'). By these means he remains convinced of his own view.

4. Distinguishing Doxastic Stubbornness and Self-Deception

What, then, is the difference between the stubborn believer and the self-deceiver, who may both unwarrantedly believe some proposition, *p*, because of a biasing desire/aversion? It seems to reside in the *sort* of desire/aversion motivating the behaviour that sustains the unwarranted belief. In cases of stubborn belief, this desire is not essentially connected to the content of the proposition believed. It will be a desire, or aversion, of a more general sort, which could make

one stubbornly cling to beliefs of varying content. For instance, Flanagan ends up stubbornly retaining his belief that Napoleon was a warmonger, but the affective factor motivating his biased thinking wasn't the desire that Napoleon was a warmonger; rather, it was the desire not to be outdone in the argument by Burke. This desire is not essentially or thematically connected to the proposition believed, and could have caused him to stubbornly retain a belief with a different content had another topic been under discussion in the same circumstances. Likewise, Burke could be on holidays with his wife and get into arguments with different people who criticise his decisions/behaviour, or point out his mistakes, for instance, with his wife about whose fault it was that they left the hotel key behind, and with another driver over a minor car accident. In all these cases he could end up stubbornly refusing to acknowledge his own culpability, to appreciate the other person's argument, or to see things from their perspective, because of the same affective factor: a strong general aversion towards being exposed as being mistaken or to blame, or towards losing an argument. This loose relationship between the biasing desire and the belief seems characteristic of stubborn belief.

Compare this to typical cases of self-deception which one finds in the (deflationist) literature. Mele talks about stock examples of ''people who falsely believe—in the face of strong evidence to the contrary—that their spouses are not having affairs, or that their children are not using illicit drugs, or that they themselves are not seriously ill'' (1997: 92). In these cases it is the desire for the proposition to be true which they unwarrantedly believe that is motivating the relevant behaviours. Jones, for instance, might explain away disturbing evidence of his wife's infidelity *motivated by the desire that she hasn't betrayed him,* and then end up believing that she hasn't. Here, the motivating desire is linked to the content of the irrational belief (indeed, they share the same content), and knowing the belief's content is crucial for understanding why he behaves as he does. This more 'specific' motivational factor that is generating the bias here, namely, the desire that his wife hasn't betrayed him, would only dispose Jones to be biased with respect to that particular issue, and not with respect to beliefs of various different contents such as with the affective factors operating in doxastic stubbornness. In cases of self-deception, the motivating desire will be of this more specific sort, and will in standard cases be the desire that *p* (where *p* is the irrationally believed proposition), or perhaps some other more specific desire in non-standard cases closely linked to the content of the belief. This distinguishes these cases from

those of stubborn belief where the subject's intransigence is motivated by a desire of a more general sort, which is more loosely related to the proposition believed.

It might be objected, however, that there are cases where people are disposed to be stubborn with respect to only a particular issue, and that in these cases, it cannot be said that the content of the desire and belief are thematically unconnected, or that knowing the belief's content is unnecessary for understanding what is motivating the stubbornness. Suppose that Flanagan is regarded as the world's leading expert on Napoleon, and has made a career out of arguing that he was one of Europe's greatest tyrants. Upstart researcher Burke, in a masterful new study reevaluating Napoleon, provides cogent reasons for thinking that Flanagan's verdict is grossly unfair. Flanagan, who has become comfortable with his status as having no equals when it comes to Napoleon, is distressed at the thought of being upstaged by the young Burke on his central area of expertise, and having his thesis refuted which is basically his life's work. Because of this, he stubbornly refuses to see the cogency in Burke's arguments and maintains confidence in his own position (to some unwarranted degree at least).

Here it may look as though knowing the content of the belief is crucial to understanding why the subject is being stubborn and unyielding with regard to it. Flanagan is particularly attached to a certain proposition about Napoleon, and is disposed to be stubborn about that, and we may even suppose, only about that. But I think that even in this case the loose relationship between the desire and belief holds true. We could explain to someone, and make it intelligible to her, why Flanagan is being stubborn in this case without making any reference to what the belief is. For we understand enough if we are told that there is a belief that Flanagan is stubbornly clinging to because it is one which he has spent his career defending and has staked his professional reputation on. Thus we can describe the culpable desire/aversion motivating his stubbornness while eliminating any reference to what the belief is about: as, for instance, the desire to be right about matters which one has staked one's professional reputation on and spent one's career defending, or the aversion to having one's life's work invalidated. That the matter concerns Napoleon in this case is a relatively incidental point, and in another close possible world, where Flanagan specialized in something else, this same desire could have caused him to stubbornly cling to a belief with a different content. Note, however, that the motivational factors presumed to be operating in Mele's stock examples of self-deception can't be described in a way

which eliminates reference to what the self-deceptive belief is about (the desire to not have been betrayed by one's wife, to not be seriously ill, or for one's children not to be taking drugs).

I don't doubt, however, that there will be cases concerning which it will be unclear whether they exemplify stubborn belief or self-deception. For one thing, the subject who falsely believes that *p* may desire that *p* and also have other desires that we would associate with stubbornness. It may not be evident which is motivating the problematic behaviour, or both may be to some extent.

5. Significance of this Phenomenon for the Motivated Bias Account of Self-Deception

In light of the above, it seems that Mele's conditions would have to be reformulated. The problem is with condition 2. In stating that the subject treats the relevant data in a motivationally biased way, this condition includes cases of stubborn belief. In fact, in a formulation of his sufficient conditions given earlier in his career (1983: 370), Mele's account didn't have this problem. According to this, *S* is self-deceived in believing that *p* when:

1. The belief that *p* which *S* has is false.
2. *S*'s desiring that *p* leads *S* to manipulate (i.e., to treat inappropriately) a datum or data relevant, or at least seemingly relevant, to the truth value of *p*.
3. This biased treatment is a nondeviant cause of *S*'s having the belief that *p*.
4. The body of data possessed by *S* at the time provides greater warrant for not-*p* than for *p*.

I speculate that the reason why Mele reformulated condition 2 is that he came to want to expand his conditions to include cases of 'twisted self-deception', which he began to dwell on afterwards. In these 'twisted' cases, subjects also believe that *p* unwarrantedly, but here they desire *that not-p*, so the desire that *p* is not motivating the bias in these cases. But in the attempt to expand his account to include these cases he has included too much else besides.

Stubborn belief shows us that, seemingly, it's not true that there are no restrictions on the kind of motivational state that can be involved in self-deception. It would be a challenging task, and one which I won't have the space to embark on here, to give an account which captures the full-range of motivational states which can be operative in self-deception, and only in self-

deception, though I may have given some pointers for an answer to this question here. To do this, one would have to consider and establish what the motivating states are in numerous sorts of cases of self-deception, 'straight' cases, 'twisted' cases, and perhaps others, while bearing in mind the sorts of desires operating in kindred but distinct phenomena like stubborn belief. However, as far as simply coming up with a set of sufficient conditions for self-deception is concerned, Mele's earlier attempt is preferable.

Acknowledgments. Thanks to Matt Soteriou, and the two anonymous referees from this journal, for their helpful comments on previous drafts of this paper. Thanks also to the audience at the European Society for Philosophy and Psychology meeting, Senate House, London, September 2012, for their comments on this material.

References

Gardiner, P. (1969–1970). 'Error, Faith and Self-Deception', *Proceedings of the Aristotelian Society*, 70, 221–243.

Holton, R. (2001). 'What is the Role of the Self in Self-Deception?', *Proceedings of the Aristotelian Society*, 101, 53–69.

Lynch, K. (2012). 'On the ''Tension'' Inherent in Self-Deception', *Philosophical Psychology*, 25, 433–450.

Mele, A. R. (1983). 'Self-Deception', *The Philosophical Quarterly*, 33, 365–377.

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

Mele, A.R. (2001). *Self-Deception Unmasked*, Princeton: Princeton University Press.

Szabados, B. (1973). 'Wishful thinking and self-deception', *Analysis*, 33, 201–205.

Szabados, B. (1974). Self-deception. *Canadian Journal of Philosophy*, 4, 51–68.