Pre-print of paper published in (2012) *Philosophical Psychology*, 25(3), pp.433-450.

On the "Tension" Inherent in Self-Deception

Kevin Lynch

Abstract

Alfred Mele's deflationary account of self-deception has frequently been criticised for being unable to explain the ''tension'' inherent in self-deception. These critics maintain that rival theories can better account for this tension, such as theories which suppose self-deceivers to have contradictory beliefs. However, there are two ways in which the tension idea has been understood. In this article, it is argued that on one such understanding, Mele's deflationism can account for this tension better than its rivals, but only if we reconceptualize the self-deceiver's attitude in terms of unwarranted degrees of conviction rather than unwarranted belief. This new way of viewing the self-deceiver's attitude will be informed by observations on experimental work done on the biasing influence of desire on belief, which suggests that self-deceivers don't manage to fully convince themselves of what they want to be true. On another way in which this tension has been understood, this account would not manage so well, since on this understanding the self-deceiver is best interpreted as knowing, but wishing to avoid, the truth. However, it is argued that we are under no obligation to account for this since it is a characteristic of a different phenomenon than self-deception, namely, escapism.

1. Alfred Mele's deflationist theory of self-deception

Alfred Mele's deflationary analysis of self-deception is one of the most influential and widely discussed accounts of the phenomenon. The elements of this analysis are not supposed to give necessary and sufficient conditions for self-deception, but only jointly sufficient conditions. Let us nevertheless call it an 'analysis'. It has, however, met with a number of objections precisely on this point; where it is argued that the conditions he lays down are *not* sufficient for self-deception.

Two of the most prominent claims to this effect are that, first, it does not include an essential intentionality condition whereby deceiving oneself is something that the subject does intentionally, or secondly, that it fails to capture the ‘‘tension’’ supposedly inherent in self-deception (Audi, 1997, p. 104; Bach, 1997, p. 105; da Costa & French, 1990, pp. 182–183; Funkhouser, 2005, p. 299; Nelkin, 2002, p. 391). In this article I will focus on the latter point, and will defend an account of this sort against this objection. However, there are two different ways of understanding this tension. On one understanding, it will be argued that a deflationist theory of this sort can indeed account for it, and in a more satisfactory way than rival accounts. But for it to do this it should be amended in a certain way. These changes will be informed by some observations on experimental work done on the biasing influence of desire on belief, work from which Mele's approach has taken inspiration. It will involve, as we shall see, switching from speaking in terms of believing or not believing a proposition, to speaking in terms of a subject's degree of conviction in a proposition. Afterwards, we will look at an alternative way of understanding the tension idea, and it will be argued that we are not obliged to account for this kind of tension, since it is characteristic of escapism rather than self-deception.

To begin, let me state the elements of the deflationary analysis, as set forth by Mele, which again are only meant to give conditions sufficient for self-deception (and which apparently are meant to capture the paradigm or ‘‘garden-variety’’ cases). The following, then, are his sufficient conditions for being self-deceived in acquiring the belief that *p*.

(1) The belief that *p* which *S* acquires 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 acquiring 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, p. 120).

Note that by ‘‘motivationally biased way’’ in (2), Mele means that *S* treats the data in a biased way, and his doing so is caused, in an appropriate way, by his having a desire/preference

regarding whether *p* is true or not. In the standard cases, which we will focus on here, *S* desires that *p*, and this causes him to be biased. Usually after giving this analysis, Mele mentions what he calls his ‘‘impartial-observer test,’’ which, as I see it, is a test for establishing whether *S*’s judgment has been biased by his/her desire.

> Given that *S* acquires a belief that *p* and *D* is the collection of relevant data readily available to *S* during the process of belief-acquisition, if *D* were made readily available to *S*’s impartial cognitive peers and they were to engage in at least as much reflection on the issue as *S* does and at least a moderate amount of reflection, those who conclude that *p* is false would significantly outnumber those who conclude that *p* is true. (Mele, 2007, p. 167)

It seems that in this test, the average judgment that the ‘‘impartial cognitive peers’’ (henceforth ‘ICPs’) make on the basis of *D* is taken as a guide to determining the judgment that *D warrants*. If the ICPs mostly judge that not-*p* on the basis of *D*, while *S* judges that *p*, then this is good evidence that *D* warrants the judgment that not-*p* and that *S*’s judgment that *p* is therefore unwarranted and also biased in favor of what she wants to be true. Furthermore, if the only relevant difference between *S* and her ICPs is that *S* desired that *p* while the ICPs lacked any strong preference regarding whether *p*, then the deviation in *S*’s judgment from what is warranted can be put down to her having that desire. On Mele’s account, assuming that *p* is false and that the causal chain between *S*’s desire and belief was non-deviant, this is sufficient to render *S* self-deceived.

Let me quickly illustrate this analysis with an imaginary case. Say that Burke is a head of state who has been accused of mismanagement, and of having made decisions responsible for putting the country into a crisis (social, economic, or military). Burke wants it to be true that his decisions were not responsible for the crisis. His having this desire causes him to reason about and evaluate the issue in a biased way. This biased thinking leads him to conclude that other factors rather than his own management and decisions are to blame (the legacy of the previous government, international forces beyond his control, etc.). However, this belief is in fact false, and also unwarranted by the evidence in his possession and available to him. It is unwarranted since if a group of Burke’s peers were to consider the issue, who have no

personal stake in the matter, are of roughly equal intelligence, and with access to the same information, they would generally conclude that he was largely responsible for the crisis. Burke's belief is therefore biased. Going by Mele's analysis, Burke is thereby self-deceived in believing that he is not responsible for the crisis.

2. The Idea of ''Tension''

Philosophers who emphasise that tension is a crucial characteristic of self-deception, however, would tend to reject this case as inauthentic, or they would at least claim that we haven't been given enough details to determine whether it's authentic. But what exactly does this tension amount to? Funkhouser divides it up into ''cognitive'' and ''behavioral'' tension (2005, p. 296). ''Behavioral tension'' may refer to the alleged fact that the self-deceiver displays some behavior that seems more consistent with believing that *p* and other behavior more consistent with believing that not-*p*. Besides this, self-deceivers supposedly also experience mental/cognitive ''conflict,'' ''tension,'' or ''discomfort'' (Funkhouser, 2005, p. 299; Graham, 1986, p. 226; Losonsky, 1997, p. 122). Graham elaborates this as the experience of being afflicted with ''doubts, qualms, suspicions, misgivings, and the like'' (1986, p. 226) concerning the belief we are self-deceived in holding, or in Losonsky's words, ''recurring and nagging doubt'' (1997, p. 122; see also Funkhouser, 2005, p. 299). Noordhof (2009) similarly speaks of an essential ''instability'' present in the self-deceived state. I would assume that these philosophers think that such mental tension is the experiential accompaniment for those cases in which behavioral tension is present or liable to occur.

Thinkers who stress the idea that self-deception involves such tension have been led to account for it in ways that are incompatible with deflationism. In particular, there are those who regard it as giving grounds for attributing contradictory beliefs to the self-deceiver. Graham says that ''the supposition that self-deception requires [believing *p* and not-*p*] can help to account for the discomfort of self-deceivers . . . . With [this], discomfort can be expected'' (1986, p. 228), and others have also used this supposition to explain this behavioral and mental tension (da Costa & French, 1990, p. 183; Demos, 1960, pp. 591–592; McLaughlin, 1988, p. 51; Scott-Kakures, 1996, pp. 48–49; Steffen, 1986, pp. 132–133). This would be incompatible with deflationism, however; Mele put forward his deflationist theory

partly as an alternative to understanding the phenomenon in terms of prima facie paradoxical features such as simultaneously held contradictory beliefs. Others still have suggested that in light of such tension, there may be no determinate answer to the question of what the person believes in self-deception (e.g., Hamilton, 2000, p. 25).

I take it that we have identified in the above remarks, under the heading of ''tension,'' a fairly definite phenomenon. It involves, on the behavioral side of things, being inclined to act in some ways that seem more consistent with believing that *p* and in others that seem more consistent with believing that not-*p*. And on the connected, mental side of things, it involves being afflicted by ''doubts, qualms, suspicions, misgivings, and the like,'' or ''recurring and nagging doubt.'' So we will take the objection to be that tension, in this sense, is an important feature of the self-deceived condition, and that a deflationary theory of the sort that Mele recommends, which rejects the contradictory belief feature, cannot account for it. My strategy for meeting this objection will not be to try to argue against the idea that such tension is characteristic of self-deception, but to affirm that it is and to modify the deflationist theory in such a way as to account for it. This modification will derive from the consideration, in the next section, of some relevant empirical work. Later, I will describe another way in which the tension idea has been understood, and my strategy here will be to argue that tension in this sense is not characteristic of self-deception, and hence is not something we are obligated to explain.

3. Experimental Demonstrations of ''Motivated Belief''

Mele's deflationary approach draws significant inspiration from experimental work on the biasing influence of desire/preference on attitudes. I will suggest that the true interpretation of the tension intuition can be found in these experiments, though this will force us to amend his approach.

Mele mentions a well-known study by Kunda (1987, third experiment), which he apparently presumes to exemplify the phenomenon he tries to capture with his sufficient conditions (Mele, 2001, pp. 11–12). In this experiment, a large group of undergraduate students read an article reviewing recent research which alleged that women who were moderate to heavy caffeine consumers were putting themselves at serious risk of developing

fibrocystic disease, said to be associated in its advanced stages with breast cancer. Female high caffeine consumers—defined as those who drank three or more cups of coffee a day—were said to be at serious risk. Subjects later had to fill in a questionnaire which included questions on how convincing they found the article to be. It was found that the female subjects who were high caffeine consumers were more skeptical of the article than were male high and low caffeine consumers and female low consumers, all of whom found the article about equally convincing.

Though he does not explicitly say so, Mele appears to treat this study as a probable demonstration of self-deception. But do any subjects in this study meet his criteria?

We can distinguish between two groups in the experiment. First, there are those who presumably felt something to be at stake for them personally in relation to the alleged link: the female high caffeine consumers. Call them the ''stakeholders.'' Then there were the rest for whom there was nothing at stake for them personally in relation to the allegation, or who were at least relatively disinterested compared to the stakeholders (some may have had loved ones who drink lots of coffee, etc.). Call these the ''non-stakeholders.'' Stakeholders, then, were found to be more skeptical of the article than non-stakeholders.

In fact, the logic of this experiment appears to reflect Mele's impartial-observer test. Non-stakeholders approximated to being Mele's ICPs. They were ''impartial'' in that they had no or less of a personal stake in the issue. They were ''cognitive peers'' in that they judged on the basis of the same body of data as stakeholders (i.e., the article), and also in that for at least some of them, we can presume there were no significant differences in relevant background beliefs (and cognitive ability) compared to stakeholders.[1] Kunda ensured this by having the alleged ill effects apply only to women, who presumably had the same prior beliefs about caffeine as the male heavy caffeine consumers in the group. The judgments of these males also differed from the stakeholders, to the same degree as the other non-stakeholders. However, closer examination of the study reveals that stakeholders fall short of passing the test in an important respect.

In the questionnaire, Kunda did not ask subjects whether they believed or disbelieved the proposition that caffeine is linked to fibrocystic disease. She asked them, rather, to indicate on a 6-point scale how convinced they were of the purported link, where ''1'' meant ''not at all convinced'' and ''6'' meant ''extremely convinced.'' The

difference between the attitudes of stakeholders and non- stakeholders, though statistically significant, was not gaping. Stakeholders' level of conviction averaged at about 3, while for non-stakeholders it averaged at the 3.5 mark. In a later replication of this study, on a 9-point scale stakeholders averaged at 5.6 and non-stakeholders at 6.72 (Liberman & Chaiken, 1992, p. 674). It thus appears that the difference between the judgments of stakeholders and non-stakeholders was subtle, and is not one where the former believed that *p* while the latter believed not-*p*.

Does this mean that the stakeholders do not pass Mele's test for self-deception? Perhaps so, though they would have to be looked at on an individual basis in line with the test (Kunda's article, however, only supplies data on them as a group). However, this only highlights the point that Mele's test may be too stringent. For consider a case where the difference in attitudes between a stakeholder, *S*, and her ICPs is described in terms of their degree of conviction. Say that on a scale from 0-10 her ICPs would mark 9 (which might mean they are ''fairly sure'' that not-*p*) while she would mark 5 (meaning, perhaps, that she's ''rather uncertain'' that not-*p*). Though this is not a case of her believing that *p* while her ICPs believe that not-*p*, her attitude towards the proposition would nevertheless have displayed a noteworthy, desire-driven deviation from what's warranted by the evidence, and this, I propose, should still attract a charge of self-deception. (In fairness to Mele, nothing he says is incompatible with this point, since he was only trying to establish sufficient conditions for self-deception when he proposed this test). If we are prepared to accept that believing that *p* when the evidence warrants the belief that not-*p*, where this is caused by a desire in the appropriate way, is sufficient for being self-deceived, then we should be prepared to accept that having an *unwarranted degree of confidence* in a proposition, where that is caused by a desire in the same appropriate way, should also qualify one as self-deceived. For this would be just a less extreme variety of essentially the same phenomenon.

4. The Notion of Degrees of Conviction

Let me say a little about this notion of degrees of conviction which I see as being in play here. Again, the psychological studies looked at above suggested this notion to us through using questionnaires on which subjects had to indicate how convinced they were of the relevant

proposition. Moreover, this seems to have been a meaningful question to have asked them. Subjects apparently understood and followed these instructions, and reliable results ensued where stakeholders were found to have marked on average lower down the scale then non-stakeholders, as was expected. These facts suggest that on some occasions there is a more fine-grained way of capturing the attitudes of subjects towards propositions than can be accomplished with the coarser conceptual apparatus of outright belief.

The notion of degrees of conviction/confidence (often called ''degrees of belief''[2]) is one that, according to Eriksson and Hájek (2007), has resisted philosophical analysis. That in itself is no reason to reject the notion, however (many patently legitimate notions defy such attempts). Like those authors, I will take it as a datum that there are such things, represented as they are in such everyday locutions as when we claim to be or feel fully convinced or certain, very convinced, fairly convinced, not very convinced, not at all convinced, etc., that *p* (where we sometimes use ''confident'' or ''sure'' instead of ''convinced'').

It is also possible to try to represent our degree of conviction numerically, which allows for much finer discriminations than would be possible with the more colloquial categories of ''fairly,'' ''very,'' etc. It is, however, important to recognize that often such scope for fine differentiation will be superfluous, and there will be indeterminacy concerning the question of what the appropriate numerical value should be to express our degree of conviction. For instance, though I might claim to feel fairly confident that a certain team will win a football game, I might not be able to decide on whether my confidence would be best represented by a 0.7 or a 0.8 on a scale going from 0 to 1, much less decide it to the second decimal place. Furthermore, it might be that no ''indirect'' test could establish which would be more appropriate, for there simply may be no fact of the matter as to whether 0.7 or 0.8 would capture my degree of conviction more correctly. However, this would not mean that I have no degree of conviction here at all, for I might not hesitate in saying that a 0.8 captures it better than a 0.2. It would only mean that degrees of conviction can't always be finely and precisely differentiated in that way. It is perhaps because of this that psychologists typically use scales of less than 10 points when measuring attitudes.

Besides relying on avowals as a measure of a subject's degree of conviction, we should also note that how convinced one is of a proposition will have implications for how much one is willing to risk on the assumption that it's true. Consequently, one's behavior in

circumstances where there are things to be gained or lost from acting on that assumption is another important measure of one's degree of conviction. (I will discuss the connection between belief and risk-taking some more later.)

There are many important philosophical questions about degrees of conviction. For instance, there is the important question of how degrees of conviction talk relates to belief talk. Perhaps the most common view here is that the concept of believing that *p* is that of having a high degree of confidence in *p*, with some arguing that the concept of belief is vague on what counts as sufficiently high (Foley, 2009; Hunter, 1996). There are also questions on how degree of conviction talk relates to talk of belief in how likely the proposition is, and on the place this notion has in the explanation of action. Luckily, we should be able to get by here without having the answers to all of these questions. I hope, therefore, to have said enough to give sufficient clarification to the notion of degrees of conviction/confidence for current purposes. Let us now turn to the issue of what use this notion can be in helping us to explain the tension inherent in self-deception, and of how it can be integrated into the deflationary account.

5. The Payoff with Changing from Talk of Belief, to Talk of Degrees of Conviction

Rethinking self-deception in terms of degrees of conviction would bring a number of advantages. Firstly, this more discriminating descriptive vocabulary allows us to talk about cases where a subject's attitude has not deviated from that of her ICPs such that the former believes that *p* while the latter believes not-*p*, but in a way that is substantial nevertheless, and that intuitively permits a charge of self-deception. It may be important for us to be able to do this, because if the above psychology experiments—which are fairly representative in this respect—are anything to go by, then deviations in confidence level between self-deceiver and ICP of a limited magnitude may be the norm when such desires bias belief. Imaginary cases which turn up in the philosophical literature where the self-deceiver believes that *p* outright while his ICPs believe that not-*p*, while heuristically useful, promote more exaggerated ideas about our powers for self-deception than would be suggested by these results. Wholeheartedly believing what one wants to be true may be rare in self-deception.

Secondly, rethinking self-deception in these terms allows us to account for the tenacious conviction that self-deception involves tension, and to do so without appropriating the controversial notion of contradictory beliefs. If we think only in terms of outright belief, the hypothesis of contradictory beliefs may seem like the best explanatory option, since neither the supposition that the self-deceiver believes exclusively that *p* or exclusively that not-*p* seems capable of explaining this tension. However, the inclusion of scales that probe confidence levels in the above studies suggests a different reason for why self-deception involves tension. Self-deceivers simply may not manage to fully convince themselves of what they want to be true. As Kunda remarks, their concerted efforts to construct justifications for their preferred positions are constrained by considerations of plausibility (1990, pp. 482–483).

In a state where one's confidence level ranges between wholehearted belief and disbelief, it would be natural to expect the aforementioned tensions to appear. Firstly, on the behavioral side, it is natural for people to make allowances for the possibility that *p*, and also to make allowances for the possibility that not-*p*, when they are uncertain as to whether *p*. For instance, to work with a type of case often mentioned in the literature, if one felt uncertain as to whether one would succumb to one's illness (and one might be self-deceived in so feeling, if the relevant evidence suggested this was an inevitability), then it would be natural for one to make allowances for what one sees as two live possibilities. One might, for instance, draw up a detailed will on the chance that not-*p*, though one might also book a holiday for the summer on the chance that *p*. Such ''ambiguous'' behavior may be in fact rational for this person in light of what she thinks is likely, though she may nevertheless be *irrational*, and self-deceived, in how likely she sees her chances of recovery as being. So if the behavioral tension of self-deception is to be understood in terms of the self-deceiver acting at times as if on the assumption that *p*, and at times also as if on the assumption that not-*p*, it can be accounted for in terms of unwarranted degrees of conviction in a proposition without need for the supposition of contradictory beliefs (though, as we will see, some philosophers may want to understand the behavioral tension idea differently).

With regards to one's phenomenology, we could here expect ''mental tension'' to arise too, if this is to be understood along the lines previously mentioned (''doubts, qualms, suspicions, misgivings, and the like,'' ''recurring and nagging doubt,'' etc.). It is true that

uncertainty in itself does not cause mental tension. Many propositions we are uncertain about give rise to no such experience, but the idea here is that uncertainty *combined with* the fact that one has a *stake* in the issue makes for the difference between merely having doubts about something, and *feeling plagued* or *nagged* by those doubts. The self-deceiver struggles to justify and find reasons for her favored position through her biased thinking, but she does not entirely succeed in countering the unwelcome evidence to her own satisfaction. Because of the stake she has in whether *p*, her doubts as to whether *p* are a source of *worry* for her, which they would not be for someone with those same doubts but without such a stake. They plague or nag her, but not a non-stakeholder in the same doxastic position.

Mele's portrayal of the self-deceiver, on the other hand, represents him as believing outright the false, unwarranted proposition, and this has led to the perception among his critics that the mental state he associates with self-deception is tension-free.[3] The term ''delusion'' is frequently reserved for this believing without tension against the evidence, presumably because the stability and surefootedness of this tension-free belief would seem to indicate an insensitivity to reason that's more the mark of pathology (da Costa & French, 1990; Funkhouser, 2005; Graham, 1986). However, a closer look at the psychology experiments from which deflationism receives its inspiration sheds some light on this issue. They suggest that deviations in confidence level between self-deceiver and ICP of a limited magnitude may be the norm when desire biases belief, and it is this fact that explains the tension of self-deception, not the alleged fact that self-deceivers hold contradictory beliefs concurrently.

I would now suggest that a set of conditions like Mele's should be reformulated in the following way. *S* is self-deceived where:

(1) *S* desires that *p*, and encounters evidence/considerations that challenge, to some extent, the assumption that *p*.

(2) The desire that *p* causes *S* to treat this evidence (reason about it and evaluate it), and to search for further evidence, in a biased way.

(3) This biased treatment (non-deviantly) causes *S* to have a degree of conviction in the proposition that *p* greater than what is warranted (by the evidence which S possesses, and which was easily available to *S* to possess), to a noteworthy extent.

Let me make a few clarifying remarks on this set of conditions. The idea of ‘‘encountering evidence’’ in condition (1) should here be taken to imply that *S* comes across the evidence and *appreciates* that it is something that, on the face of it, poses a challenge or threat to the assumption that *p*. This is necessary to motivate the attempts to deal with it (by, for instance, trying to explain it away) referred to in condition (2), though this appreciating shouldn't imply that *S* draws the conclusion that this evidence warrants.

Moreover, the notion of unwarrantedness here is to be understood in relation to the degree of conviction of *S*'s average IPC. *S*'s degree of conviction in the proposition that *p* will be unwarranted if it deviates to a noteworthy degree from that which her ICPs would form on the basis of considering the same information that *S* was acquainted with, and deviates in the direction of what *S* wants to be true. Note also that I take it, as I think Mele does too, that we should think of an attitude as being unwarranted relative to, not just the evidence that *S* possesses, but possibly also to evidence which *S* didn't possess but which *S* easily could or should have possessed, i.e., evidence (or considerations) which were ‘‘easily available’’ to her. This is because the biased behavior referred to in (2) may include selective gathering of further evidence, and in such cases we should consider the degree of conviction as being unwarranted relative to the evidence/considerations that she *neglected* to collect/consider, because of this selective evidence search.

Note further that unlike Mele, I am omitting any condition which states that *p* must be false, since I don't think it needs to be false. For a start, this would rule out our speaking of any stakeholders in experiments like that of Kunda's as being self-deceived, where the welcome proposition for all we know is true, with the apparently credible evidence against it being a mere fabrication of the experimenters, and this seems like an unwelcome result. But to address Mele's basis for including this condition more directly: although Mele may be correct in saying that ‘‘*S* is (self-) deceived in believing that *p*’’ entails that *p* is false, ‘‘*S* deceived herself into believing that *p*’’ does not, and we can use this construction instead when attributing self-deception when *p* happens to be true. I deal more fully with this issue elsewhere (Lynch, 2010).

Before leaving this section, let me attempt a general explanation of why self-deception involves tension of this sort. Self-deception is a phenomenon of normal, not abnormal, psychology. It is, by definition, not a pathological phenomenon, as delusion is.

Consequently, it is something that is perpetrated by normal people. It is partly constitutive in turn of the idea of a normal person that they are, in general, intellectually able and rational, and are consequently *not* completely immune to the force of good evidence when they encounter it. For insofar as someone showed himself to be insensitive to reason entirely, to that extent his condition would be considered abnormal and pathological. So it is the fact that self-deception is, by definition, perpetrated by normal people, who are generally sensitive to the force of good evidence, that explains why it must be characteristic of self-deceivers not to be left entirely unperturbed by the weighty evidence, try as they might to dismiss and ignore it. And it is this which also explains why motivationally biased belief only occurs, as numerous writers have remarked, when the unwelcome evidence falls short of being conclusive, and why it ''evaporates when exposed to the light of overwhelming fact'' (Johnson, 1997, p. 118).

Does this mean that it is a conceptual or empirical truth that self-deception involves tension? It is not clear what to say here. On the one hand, we may want to agree with Mele in saying that it is not conceptually *necessary* that it does (Mele, 1997a, p. 131). For may it not be possible for self-deceivers on occasion to be entirely successful in explaining away the unwelcome evidence through their biased behavior to their own satisfaction, though this would not make them count as delusional in a pathological sense since they would still be disposed to assent to the unwelcome truth were incontrovertible evidence produced? I can see no reason to deny that this could ever occur, or to claim that if it did occur we should not call them self-deceivers. On the other hand, tension is associated with self-deception because self-deceivers are generally rational beings who are generally sensitive to the force of good evidence, and that is not a contingent truth about self-deceivers. If conceptual truths are all necessary truths then perhaps it's not a conceptual truth that self-deception involves tension, but this may be too narrow a view of what conceptual truths are.

6. Deep Conflict Cases

If this is what the tension of self-deception amounts to—feeling harassed by doubts, mental wavering, and the behavior associated with such uncertainty, etc.—then the above modified version of Mele's deflationism is well suited to account for it. It is not clear, however, that this account would manage so well with other ways in which the idea of tension has been

understood. For some philosophers understand this tension in a way that, as I will argue, does not seem compatible with the idea that the self-deceiver is overly confident in the welcome proposition, relative to the confidence that their evidence warrants. They would probably accuse me of having mischaracterized the explanandum above.

Some philosophers understand the behavioral tension of self-deception in a particular way, where they imagine the self-deceiver as indicating with *what she says* that she believes that *p* (perhaps by denying that not-*p* in a defensive, evasive, or flustered fashion), and as indicating with *non-verbal behavior* that she knows the unwelcome truth, not-*p* (e.g., Audi, 1997; Lee, 2002, p. 282). The relevant non-verbal behavior most often mentioned is *avoidance behavior*, in which the self-deceiver steers clear of things that may remind her of, or put her face-to-face with the unwelcome truth, behavior which points strongly to belief in the truth (Funkhouser, 2005; Patten, 2003, p. 241; Pears, 1991, p. 398; Williams, 1970/1993, p. 151). Funkhouser (2005) illustrates the structure of a supposedly typical case with two examples. First is the example of a balding man who denies that he's bald and yet ensures that his baldness is kept hidden (using the ''comb-over technique,'' posing at a certain angle for photographs, refusing to let his wife tussle his hair, etc.). The second example is of a woman called Nicole who avows to her concerned friends and to herself that her husband is not having an affair with a certain other woman (in the face of strong evidence to the contrary), but who goes out of her way to avoid places where she would find them together if the reports were true. Funkhouser (2009) calls these cases of ''deeply conflicted'' self-deception. I will adopt this terminology, but without assuming that these are cases of self-deception, by calling them ''deep conflict'' cases.

Regarding these deep conflict cases, what should we conclude from such behavior about the person's beliefs? Some deep conflict theorists have thought that this would be evidence that the subject both believes that *p* and believes that not-*p*. Where *p* is the welcome falsehood, they have thought that their verbal assertions that *p* are evidence that they believe that *p*, while the avoidance behavior would be evidence that they also believe not-*p* (e.g., Rey, 1988, pp. 264 & 278). Others have suggested that there may be no determinate answer to the question of what the person believes (Funkhouser, 2009; Hamilton, 2000, p. 25). These responses seem motivated by the assumption that the relevant belief-that-*p*-consistent and belief-that-not-*p*-consistent behavior are of *equal evidential weight* as indicators of what the

subject believes. This is questionable. As Funkhouser (2005, p. 300) notes, we generally take non-verbal behavior to be privileged over verbal behavior when it comes to belief attribution, and he takes (in his earlier paper, though not his later one) these deep conflict cases as being indicative of someone who knows the truth and who does not really believe the contrary falsehood at all, despite what he says. Funkhouser seems to me to be correct in this, though he wonders why we privilege non-verbal behavior. The explanation for this, I contend, can be given in terms of risk-taking, an explanation hinted at by Funkhouser (2005, p. 307) himself, and developed by Gendler (2007).

7. Belief and Risk-Taking

It is not true that all belief-consistent behavior is of equal weight for deciding on what someone believes. What is of paramount importance is how the person acts when he/she understands that there would be something to be gained, or costs to be incurred, if one were to act on the assumption that the belief is true, were it actually not true. That is, the extent to which *S* really believes that *p* can be gauged by observing the risks he/she is willing to take on that assumption. Therefore not all belief-that-*p*-consistent behavior is of equal evidential worth, because not all belief- that-*p*-consistent behavior is associated with equal levels of risk or of possible gain. For instance, my act of saying that the food isn't poisonous doesn't carry the same weight as my act of tasting the food, when deciding whether I believe the food isn't poisonous, other things being equal.

This is relevant to our assessment of the subject's attitudes in the deep conflict cases purported to be cases of self-deception. Consider the following such case from Gendler (2007, pp. 244–245), where a man who has been diagnosed as terminally ill denies and gives explanations against the diagnosis, suggesting that he believes he is well (*p*). This is belief-that-*p*-consistent behavior. However, he later comes by the opportunity to take a powerful drug that would cure that illness, but that would be detrimental to anyone without that illness. He opts to take the drug. This is belief-that-not-*p*-consistent behavior. Clearly though, both behaviors don't have equal evidential weight. The reason is that his acting as if *p* were true when quizzed by others about his health would not cause him to incur any significant cost or loss if it were really the case that not-*p*. If he were pretending, it would cost nothing to keep up

the pretence in these circumstances. But his acting as if *p* were true when he's offered the drug (which would involve not taking it) would cause him to incur a significant loss were it really the case that not-*p*. If the man was pretending that *p* here, then maintaining the pretence would be very costly, in that he would miss an important opportunity. This is why the belief-that-not-*p*-consistent behavior evidentially trumps the belief-that-*p*-consistent behavior, and we reasonably infer, as Gendler recommends, that he really thought that he was terminally ill after all, and must have been just pretending to others, and to himself, that things were fine. These points account for why ''actions speak louder than words,'' that is, why we generally prioritize non-verbal over verbal behavior when ascertaining belief, and they allude to classic methods for exposing malingering. Of course, circumstances sometimes do obtain where a lot rides on verbal expressions of belief (think of someone in a game-show with big prizes at stake), but more often risk is associated with non- verbal behavior.

In deep conflict cases, as they are typically described, the subject's belief-that-*p*-consistent behavior would not be at all costly if it were true that not-*p*. It is just verbal behavior: usually just a matter of saying things in front of people. However, the subjects fail to display belief-that-*p*-consistent behavior in circumstances where it would be costly to act on that assumption, given that not-*p*. I take it that we would ordinarily consider this good evidence that the person doesn't really believe what he/she professes to believe or to be true at all. Verbal assertions that *p* do not support a judgment that one believes that *p* where there is an unwillingness to put anything on the line on that assumption. Genuinely believing that *p* implies a willingness to, so to speak, ''put one's money where one's mouth is.'' Because subjects in these cases are best interpreted, then, as knowing the truth that not-*p*, they could not be accounted for with the modified deflationary theory given above, which demands that they have a degree of confidence in the welcome proposition that *p* which seems incompatible with our saying that they know the contrary unwelcome one. I now wish to argue that with these cases, self-deception has been confused with something else.

8. Self-Deception and Escapism

The majority of philosophers have held that self-deception involves false or unwarranted belief, but the view that it may not has been around for quite some time. However, it is

worth noting that those who maintain this usually do not deny that self-deception may sometimes involve being self-deceived in what one believes, where that belief is unwarranted and typically false. Possibly the first philosopher to advance this view was Martin (1979). For Martin, self-deception ''need not involve ignorance and unwarranted belief'' (1979, p. 446), but may instead involve the ''intentional evasion of unpleasant topics and truths'' (1997, p. 122), truths which one knows to be true. This ''*need* not'' suggests that Martin thinks self-deception *can* involve ignorance and unwarranted belief. Though Funkhouser (2005) initially categorized unwarranted belief cases as ''self-delusion'' rather than self-deception, reserving the term 'self-deception' for deep conflict cases, he later renounced this, thinking that his original distinction was ad hoc and stipulative. His new view (Funkhouser, 2009) is that deep conflict cases are only *one kind* of self-deception, along with unwarranted belief (or preferably, degree of conviction) cases of the sort that Mele focuses on. Gendler takes deep conflict cases to be ''the cleanest and most interesting cases of self-deception'' (2007, p. 233), which again implies that she doesn't deny that the rival false/unwarranted belief cases count as self-deception also. Although some philosophers still deny that false/unwarranted belief cases are self-deception, preferring to call them ''delusion'' (e.g., Audi, 2007), Mele (2010) has recently done some experimental philosophy, and has presented evidence that naive subjects are inclined to use ''self-deception'' for false/unwarranted belief cases. (Though suggestive, this evidence may not be conclusive; one would want to know much more here, like how they would categorize deep conflict cases, and whether after considering those, they would be inclined to revise their initial decision with the unwarranted belief cases, thinking deep conflict cases as being more deserving of the title ''self-deception'').

So it seems that we have on our hands two distinct phenomena claimed to belong to the same psychological category. Roughly, there is (1) avoiding reflecting on and confronting an unpleasant truth that one knows about, and (2) not believing this unwelcome truth, but having an unwarranted skepticism towards it due to your desires biasing your evaluation of the issue. Now deep conflict theorists, who regard (1) as instantiating self-deception, will if pressed generally not deny the widespread presumption that (2) counts as self-deception also.

So the question, then, is whether both of these phenomena are in fact self-deception. Do (1) and (2) constitute a single psychological kind? It would be *surprising* if this were so, for a number of reasons. First, there is a world of difference between (1) and (2).

The concept of self-deception would be thus ambiguous. For if someone told me, ''Jones, he's deceiving himself about such-and-such,'' I would not know what to think, and would have to follow up, ''do you mean he doesn't believe that such-and-such is true, or do you mean he knows it but is avoiding it, won't talk about it and admit it, etc.?'' Secondly, though we can see the similarities between (2) and interpersonal deception, and can hence appreciate why they would both count as species of the one genus (i.e. deception),[4] interpersonal deception and (1) have little in common, making it a mystery why they would be considered species of the same genus.[5]

Thirdly, and most importantly, these theorists have overlooked the fact that we have another term used for picking out (1): 'escapism'. One of the few people in the self-deception literature who has shown sensitivity to the distinction between self-deception and escapism, in an article deserving of more attention than it has received, is Longeway (1990). Escapism, Longeway suggests, comes in mild and harmless forms, as when we indulge in entertainments which temporarily draw us away from our everyday troubles. But sometimes, he remarks:

> We speak of a more serious escapism, in which we avoid thinking about what we know to be so, not in the course of recreation or to keep unpleasant thoughts out of mind as long as they are not necessary, but as a defense against reality itself. (Longeway, 1990, p. 1)

The escapist ''attempts to keep beliefs one does not like out of consciousness … and should they enter consciousness, to distract one from them or put them out of mind'' (Longeway, 1990, p. 2). These activities, he claims, exercise us quite often, though they don't attract the charge of escapism unless we habitually try to avoid the reality when we should or need to consider or face up to it. Examples of escapist techniques mentioned by Longeway (1990, p. 1) include distracting oneself with irrelevant concerns to force the belief out of consciousness, denying verbally or pretending to oneself or others not to hold it, avoiding situations which would remind one of the matter, and restricting one's company to those who will not remind one of it, all of which are activities deep conflict theorists have associated with self-deception. Things used for escapist purposes can be anything from one's

own imagination, to alcohol and drugs. Longeway says that many of the techniques of escapism may be used for self-deception too, that is, for the purpose of eliminating the belief rather than avoiding awareness of it (he may be showing his intentionalist allegiances here), and that is why the two phenomena, self-deception and escapism, can easily be confused with one another. But for Longeway the crucial point that distinguishes escapism from self-deception is that the former involves avoiding reflecting on and being reminded of an unpleasant reality that one knows about, whereas the latter involves having an unwarranted belief towards the matter.

Perhaps, then, we *have* found the differences between cases (1) and (2) worth acknowledging in our conceptual scheme, and the terms 'self-deception' and 'escapism' have been supplied for this purpose. Deep conflict cases resemble paradigm cases of escapism, and thus Martin, Gendler, Funkhouser and company (see also Bach, 1981) are open to the charge of failing to respect the distinction between an escapist and a self-deceiver (though escapism and self-deception may frequently be found mixed up together in real life cases, which is one reason why they may be so easily confused[6]). Again, this all presupposes that the best way to interpret the subject's behavior in these cases is as indicating that they really know the truth. From the descriptions given of such cases in the literature, this presumption seems justified, but whether this is the best explanation of any case ultimately depends on its details and how it is described. On other descriptions, the behavior may not weigh so heavily towards that conclusion, indicating some degree of confidence in the welcome proposition (evidenced by a willingness to take some risks on the assumption that *p*, for instance), and so the case may be amenable to treatment with the above modified deflationary account if this confidence goes beyond what's warranted. The important thing is just to bear in mind what distinguishes escapism from self-deception, which is that the former involves knowing/believing (''deep down'') an unwelcome truth but avoiding facing up to it, while the latter would require that the subject is skeptical towards this proposition to some degree.

The idea being suggested, then, is that when our attention is drawn to the overlooked notion of escapism and how it is defined, and when we see that deep conflict cases resemble paradigmatic instances of escapism, we should feel less of a temptation to stretch the term ''self-deception'' to cover these kinds of case as well as the unwarranted belief

cases which so typically go under that title. It may, in addition, be pointed out that part of what motivates the thought that these deep conflict cases are self-deception may be the idea that self-deception must be closely modeled on interpersonal deception. For typically in interpersonal deception, the deceiver does know the truth, so one might expect that the self-deceiver must know the truth also. However, as has been pointed out before (Champlin, 1977), this reasoning is specious. Very briefly, the usual strategy to demonstrate this is to show that if we were to apply this reasoning to the notion of being self-taught, then the one who is self-taught would have to know beforehand that which he sets out to learn, since in the interpersonal case where one teaches another, the teacher knows beforehand what she sets out to teach the learner. This result is supposedly absurd (I deal with this issue more fully in Lynch, 2009).

9. Conclusion

In conclusion, if the claim that self-deception involves tension is to be understood as meaning that the self-deceiver is usually afflicted with nagging doubts about what he wants to be true, and exhibits behavior which indicates something short of wholehearted commitment to the welcome assumption, then this idea can easily be assimilated into a deflationary account like Mele's which frames things in terms of unwarranted degrees of conviction. This gives us a way of viewing the mind of the self-deceiver that is both logically less provocative than the alternative contradictory belief account, and more in line with empirical work on the biasing influence of desire on belief. If, however, the idea is to be understood as meaning that the self-deceiver displays behavior which weighs more in favor of the conclusion that she knows the unwelcome truth and is trying to avoid confronting it, then we are under no obligation to accommodate such a thing in our theory, since it is the mark of an evader and an escapist, and not a self-deceiver.

Acknowledgements

I'd like to thank Alfred Mele, Johannes Roessler, and the anonymous reviewers of this journal for their helpful comments on this paper or parts of it, and also to the audiences at *The*

Nature of Belief conference, University of Southern Denmark, Odense, 2010, and at seminars in Warwick University, where material from it was presented.

Notes

[1] Note that to qualify as ICPs, the stakeholders would also have to put a reasonable amount of effort into assessing the issue. The idea here is presumably that we should not put much stock into the judgments of people who assess an issue in a cursory way, as people for whom the issue does not matter could be inclined to do. Rather, an ideal judge would be one who is impartial, and yet motivated to come to an accurate judgment on the issue (having what psychologists call ‘‘accuracy motivation’’). Though in this experiment Kunda did not seem to make any special efforts to ensure that non-stakeholders were so motivated, the experimental context might have supplied this somewhat. Furthermore, other studies that have made such efforts have presented similar deviations between the judgments of stakeholders relative to non-stakeholders to what’s witnessed here (Lundgren & Prislin, 1998).

[2] This choice of expression has some disadvantages, and is one I wish to avoid. Theorists who employ this terminology often represent the range of degrees of belief on scales ranging from 0 to 1, where 0 means that one is certain that not-*p*, and where 1 means one is certain that *p*. However, many also believe that talk of believing that *p* simpliciter can be understood as having a sufficiently high degree of belief. Consequently, one can fail to believe that *p* because one has too low a degree of belief that *p*. But this is odd: one would think that having any degree of belief towards *p* presupposes that one believes that *p*, just as feeling angry towards A to some degree presupposes that one feels angry towards A, even if only a little.

[3] Mele does say that his ideas on self-deception could be formulated in terms of ‘‘degree of belief/confidence’’ (2001, p. 10). However, he doesn’t exploit this possibility for dealing with the tension issue, though he perhaps hints at this possibility when he suggests one way of accounting for behavioral tension cases by saying that the self-deceiver may believe that *p* while believing there’s a significant chance that not-*p* (1997b, p. 96).

[4] Two salient similarities are that (1) both cases typically involve the deceived having either a false or an unwarranted belief, and (2) that in both cases, the actions of the deceiver are responsible for the deceived having this problematic belief, though in self-deception, of course, the deceiver is the same person as the deceived.

[5] Funkhouser (2005, pp. 299 & 304) anticipates this objection, and responds by saying that since conceiving of self-deception on the model of the interpersonal case leads to well-known problems, we shouldn’t feel that there needs to be a close similarity between self-deception and

interpersonal deception. But surely there must be *some* similarity, *some* shared features between the two cases to sustain them both as species of deception, and deep conflict cases simply don't appear to have the requisite similarity.

[6] The alcoholic, for instance, who uses alcohol as a means of escape from an unpleasant reality, might have also deceived himself about his true motives for using alcohol, or into believing that he can indefinitely avoid the unpleasant reality, or that his present behavior is sustainable, or that he's a victim, or that he doesn't have it in him to change.

References

Audi, R. (1997). Self-Deception vs Self-Caused Deception: A Comment on Professor Mele. *Behavioral and Brain Sciences,* 20, 104.

Audi, R. (2007). Belief, Intention, and Reasons for Action. In M. Timmons, J. Greco, & A. Mele (Eds.), *Rationality and the Good* (pp. 248–262). New York: Oxford University Press.

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

Bach, K. (1997). Thinking and Believing in Self-Deception. *Behavioral and Brain Sciences,* 20, 105.

Champlin, T. S. (1977). Self-deception: A Reflexive Dilemma. *Philosophy,* 52, 281–299.

da Costa, N.C.A., & French, S. (1990). Belief, Contradiction, and the Logic of Self-Deception. *American Philosophical Quarterly*, 27, 179–197.

Demos, R. (1960). Lying to Oneself. *Journal of Philosophy,* 57, 588–595.

Eriksson, L., & Hájek, A. (2007). What are Degrees of Belief? *Studia Logica,* 86, 183–213.

Foley, R. (2009). Beliefs, Degrees of Belief, and the Lockean Thesis. In F. Huber & C. Schmidt-Petri (Eds.), *Degrees of Belief* (pp. 37–47). London: Springer.

Funkhouser, E. (2005). Do the Self-Deceived Get What They Want? *Pacific Philosophical Quarterly*, 86, 295–312.

Funkhouser, E. (2009). Self-Deception and the Limits of Folk Psychology. *Social Theory and Practice*, 35, 1–13.

Gendler, T. S. (2007). Self-Deception as Pretense. *Philosophical Perspectives,* 21, 231–258.

Graham, G. (1986). Russell's Deceptive Desires. *The Philosophical Quarterly*, 36, 223–229.

Hamilton, A. (2000). The Authority of Avowals and the Concept of Belief. *European Journal of Philosophy,* 8, 20–39.

Hunter, D. (1996). On the Relation between Categorical and Probabilistic Belief. *Noûs,* 30, 75–98.

Johnson, E. A. (1997). Real Ascriptions of Self-Deception are Fallible Moral Judgments. *Behavioral and Brain Sciences,* 20, 117-118.

Kunda, Z. (1987). Motivated inference: Self-Serving Generation and Evaluation of Causal Theories. *Journal of Personality and Social Psychology*, 53, 636–647.

Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin*, 108, 480–498.

Lee, B. D. (2002). Shoemaker on Second-Order Belief and Self-Deception. *Dialogue*, 41, 279–289.

Liberman, A., & Chaiken, S. (1992). Defensive Processing of Personally Relevant Health Messages. *Personality and Social Psychology Bulletin*, 18, 669–679.

Longeway, J.L. (1990). The Rationality of Escapism and Self-Deception. *Behavior and Philosophy*, 18, 1–19.

Losonsky, M. (1997). Self-Deceivers' Intentions and Possessions. *Behavioral and Brain Sciences,* 20, 121–122.

Lundgren, S.R., & Prislin, R. (1998). Motivated Cognitive Processing and Attitude Change. *Personality and Social Psychology Bulletin*, 24, 715–72.

Lynch, K. (2009). Prospects for an Intentionalist Theory of Self-Deception. *Abstracta*, 5, 126–138.

Lynch, K. (2010). Self-Deception, Religious Belief, and the False Belief Condition. *The Heythrop Journal*, 51, 1073–1074.

Martin, M.W. (1979). Self-Deception, Self-Pretence, and Emotional Detachment. *Mind*, 88, 441–446.

Martin, M.W. (1997). Self-Deceiving Intentions. *Behavioral and Brain Sciences,* 20, 91–102.

McLaughlin, B.P. (1988). Exploring the Possibility of Self-Deception in Belief. In B.P. McLaughlin & A.O. Rorty (Eds.), Perspectives on Self-Deception (pp. 29–62). Berkeley, CA: University of California Press.

Mele, A.R. (1997a). Understanding and Explaining Real Self-Deception. *Behavioral and Brain Sciences,* 20, 127–136.

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

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

Mele, A.R. (2007). Self-Deception and Three Psychiatric Delusions. In M. Timmons, J. Greco, & A. Mele (Eds.), *Rationality and the Good* (pp. 163–175). New York: Oxford University Press.

Mele, A.R. (2010). Approaching Self-Deception: How Robert Audi and I Part Company. *Consciousness and Cognition*, 19, 745–750.

Nelkin, D. (2002). Self-Deception, Motivation, and the Desire to Believe. *Pacific Philosophical Quarterly*, 83, 384–406.

Noordhof, P. (2009). The Essential Instability of Self-Deception. *Social Theory and Practice*, 35, 45–71.

Patten, D. (2003). How do we Deceive Ourselves? *Philosophical Psychology*, 16, 229–246.

Pears, D. (1991). Self-Deceptive Belief Formation. *Synthese*, 89, 393–405.

Rey, G. (1988). Towards a Computational Account of Akrasia and Self-Deception. In B.P. McLaughlin & A.O. Rorty (Eds.), *Perspectives on Self-Deception* (pp. 264–296). Berkeley, CA: University of California Press.

Scott-Kakures, D. (1996). Self-Deception and Internal Irrationality. *Philosophy and Phenomenological Research*, 56, 31–56.

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

Williams, B. (Ed.). (1993). Deciding to Believe. In *Problems of the Self: Philosophical Papers, 1956–1972* (pp. 136–151). London: Cambridge University Press. (Original work published 1970.)