The Essential Instability of Self-Deception

Two apparent paradoxes lie at the heart of discussion of self-deception, one focusing on belief, the other on intention. The belief paradox concerns how the self-deceived can combine the belief that p and the belief that not-p. The intention paradox concerns how the self-deceived can intend to believe that p, and manage it, without knowing what they are up to and vitiating it. Both are said to be paradoxes because, on the one hand, self-deception seems possible and, on the other, it can seem to require combinations of states that render it impossible.

The first choice point for debate is whether to divide or dilute. Dividing presses on the analogy with the deception of others. There is no problem with Jo believing that p and Josephine believing that not-p; nor is there a problem with Jo intending to bring it about that Josephine believe that not-p and managing to make it so. The division strategy seeks to appeal to this fact and relocate the division, in some attenuated sense, within subjects so that they can genuinely count as *self*-deceived. Dilution explains how the allegedly paradoxical combination of states is not required. Instead, self-deception involves something less that is not paradoxical.

Both approaches suffer from a problem—in fact, the same one. Each gets rid of the paradoxical character of self-deception at the price of losing the instability that is essential to it. The problem with the self-deceived is that they seem to avoid accepting a certain proposition and have anxiety over, or lack confidence in, what they are up to. It seems as though the project may fail or requires work. I put all this in terms that are as neutral as possible. It is pretty clear how the two paradoxes with which I began involve a more precise articulation of it. Their anxiety or lack of confidence stems from the fact that, deep down, they believe the proposition and have intentionally produced a belief in the opposite whose work will be unpicked if they appreciate what they have done.

Thus, I say, characterize the instability and the essential work of characterizing self-deception is done. The approach has a number of advantages that I seek to bring out in the course of the paper. The first is that it provides unity where other accounts of unity fail. As we shall see, there are a number of different ways people can be self-deceived (in terms of combinations of states) and it is a mistake to try to single out one form only. The second is that diluted accounts are subject to counterexamples, or have unfortunate commitments, which can be avoided if an appeal to instability is added.

The second point needs careful handling. Self-deception is plausibly seen as one kind of theoretical irrationality: irrationality about what to believe, sincerely avow, or in some other way cognitively endorse as true. To keep the options open, I shall talk of *cognitive endorsement* and understand this to cover the other two types of states just identified, and take *endorsement* to be to endorse as true. Self-deception is often contrasted with wishful thinking (in which subjects cognitively endorse a proposition because they want it to be true) and full-blown delusion (which I shall discuss more fully in section 4), but in some way involves a subject losing grip on reality with regard to a certain subject matter so that they have little chance of being able to make appropriate cognitive adjustments to the way the world is. As the American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* puts it, delusion is:

> A false belief based on incorrect inference about external reality that is firmly sustained despite what almost everybody else believes and despite what constitutes incontrovertible and obvious proof to the contrary.[1]

The careful handling to which I refer is partly due to the fact that since we are in the empirical business of identifying, presumably, sometimes instantiated mental kinds, there is no reason to suppose that our present judgments about what is involved in self-deception, in contrast to these other phenomena, will survive scrutiny. An appropriate taxonomy may point in another direction. One of the aims of this paper is to argue that an appropriate demarcation should appeal to a certain kind of instability.

Careful handling is also required because it seems that our talk of self-deception may involve two substantially different senses of deception. According to the first, *external*, way, the crucial difference is that the resultant cognitive endorsement is false. Subject are deceived by themselves because they are, in some way, responsible for the fact that they have arrived at a false cognitive endorsement. According to the second, *internal*, way, subjects are deceived by themselves because they are, in some way, responsible for the fact that they have arrived at a cognitive endorsement that, by their own lights, they take to be or suspect is false. This is compatible with the cognitive endorsement being true. Maybe

[1] Page 765, taken from Martin Davies, Max Coltheart, Robyn Langdon, and Nora Breen, "Monothematic Delusions: Towards a Two-Factor Account," *Philosophy, Psychiatry, and Psychology* 8, nos. 2-3 (2001): 133-58, p. 133.

Alfred Mele is right that we should add that the cognitive endorsement is false simply because that's what *deception* means.[2] Yet, arguably, this would lose an important dimension of commonality between true and false cases of internal self-deception.

The most significant contribution of diluting accounts is that they have rightly questioned whether all self-deception must involve the allegedly paradoxical combinations of states even if they have failed to establish that no self-deception can include the combinations in question. Indeed, the foremost proponent of a diluting account—Mele—seems to accept that at best the paradoxical combinations are empirically unjustified rather than logically or metaphysically impossible.[3] This raises the question of why attempt dilution at all. The right answer is that the diluters are onto something—as I already mentioned—but have fumbled the justification for their position. They can argue that weaker combinations are all that is required when they give rise to the essential instability of self-deception. They don't have to grit their teeth and say: "limit self-deception to that if you want to but it is not obvious that the case you have provided should be described in that way."

In section 1, I will focus on the question of whether self-deceptively supported cognitive endorsements that p are the result of a desire that p, a desire for the cognitive endorsement that p or an intention that one cognitively endorse that p. Broadly, there are two arguments for holding that it is a desire for the cognitive endorsement that p or an intention that one cognitively endorse that p. First, it is suggested that an action, or more generally purposive, explanation of the cognitive endorsement that p is not available if a subject desires that p.[4] In this respect, appeal to a desire for the cognitive endorsement that p is better. Second, it is suggested that appeal to a desire for the cognitive endorsement that p, or an intention that one cognitively endorse p, is preferable, because it presents the best chance of self-deception being a unified phenomenon.[5] Twisted cases of self-deception (in which subjects self-deceptively believe what they don't want to be true) vitiate this prospect if we take self-deception generally to involve a desire that p, since then we are forced to treat twisted cases as special.[6]

I shall argue for the following three claims. First, it is compatible with

[2] Alfred R. Mele, *Self-Deception Unmasked* (Princeton: Princeton University Press, 2001), pp. 50-51.

[3] E.g., ibid., p. 17.

[4] Dana K. Nelkin, "Self-Deception, Motivation, and the Desire to Believe," *Pacific Philosophical Quarterly* 83 (2002): 384-406, pp. 396-97; Mele, *Self-Deception Unmasked*, pp. 14, 23.

[5] Nelkin, "Self-Deception, Motivation, and the Desire to Believe," pp. 395-96.

[6] The term "twisted" is Mele's: see *Self-Deception Unmasked*, chap. 5.

self-deception involving a desire that p that an agent-style explanation can be provided of the resultant cognitive endorsement that p. Second, Mele's appeal to desires' influence on confidence levels for, specifically, believing that p or believing that not-p provides limited insight into the nature of self-deception, and hence it is not clear that an alternative to agent-style explanation has been identified. Third, it is a mistake to attempt to unify cases of self-deception by appeal to a desire for belief, or cognitive endorsement more generally.

The upshot of section 1 is that things look good for those who try to unify self-deception by appeal to agency. In section 2, I argue that this masks a diversity of kinds of agent-style explanation that may be provided. So we need to look for unity elsewhere. In section 3, I discuss how appeal to the essential instability of self-deception in attentive consciousness can provide this unity and, importantly, a come-back for Mele's approach appealing to desires' influence on confidence levels. I develop this point to illustrate further how such an appeal enables diluting accounts of various kinds to avoid counterexample, in particular, those that reject the idea that the self-deceived must both believe that p and believe that not-p. So I make good on my claim that there is unity in the face of considerable diversity. In section 4, I discuss the differences between self-deception and delusion. I compare, favorably, my instability-based account with Mele's appeal to motivational factors.

1. Agency versus Non-Agency Views of Self-Deception

Agency views of self-deception hold that self-deceptively favored cognitive endorsement is produced by some, perhaps attenuated, form of agency. The most straightforward way in which this may be understood is that such cognitive endorsements are intentionally produced. Anti-agency views deny that these cognitive endorsements are produced as a result of agency. Instead, anti-agency views take desires and, perhaps, emotions more broadly to have a direct influence upon self-deceptively favored cognitive endorsements.[7] Both agency and anti-agency views have an explanatory burden: what is the mechanism by which desires and other emotions influence subjects' cognitive endorsements? The attempt to satisfy this explanatory burden is one way to get traction on the issue of what is the most plausible attribution of states in virtue of which the product of self-deception is achieved: via an intention to cognitively endorse that p, a desire that p, or a desire for a cognitive endorsement that p.

Suppose that I want to believe that I am a good driver or want to be-

[7] The terminology is Mele's: see ibid., p. 13.

lieve that I am a good judge of character. Why do these (when they do) result in the belief that I am a good driver or the belief that I am a good judge of character? The problem is that the desires seem to concern states of the world and not beliefs we may have. Yet, the desires are supposed to result in beliefs, or cognitive endorsements more generally. We don't get what we desire and yet the desires are supposed to explain what we get.

In the case of agency accounts, one way in which there would be a connection between subjects' desire that p and belief that p is if they had the means-end belief that if they believe that p, then p is the case. Nevertheless, this would be a most peculiar means-end belief for subjects to have. It attributes a power to thought that, sadly, is rarely evident.[8] Fortunately, there is a better alternative. If we desire that p, then, in general, we also desire to believe that p. If you want the world to be a certain way, then you don't want it to be that way without also believing it to be that way. There will be exceptions—indeed we could fix up a fantasy case in which coming to believe it would destroy the very thing we want—but agency explanations will be available for all the others. S desires that she believe that p, believes that by doing such and such, she will have the belief and hence, given that S does those things, S will have the belief that p.

The legitimacy of attributing the desire for a belief that p is also indicated by the fact that the self-deceived often seem to try to avoid believing that not-p by looking for and assessing evidence in a biased way.[9] If they desired to have a belief that in fact was not supported by the evidence, then this behavior would be explained.

The move just made will seem like grist for the mill of those who emphasize that self-deception always involves a desire for a cognitive endorsement that p rather than a desire that p. However, it is important to distinguish the appropriate characterization of cases of self-deception from the explanation of how those elements relate to each other. The point I have just made is that a desire that p can be central to a characterization of self-deception and yet explain why it is also legitimate to attribute other desires that are responsible for the connection between the desire that p and the resulting cognitive endorsement that p.

Of course, if it turned out that all cases of self-deception involved a desire for cognitive endorsement that p, then those who emphasize this element would be in a strong position to argue that this should be a central feature in an analysis of self-deception. However, this is not the case. First, as we shall shortly see, desires can have an influence upon beliefs

[8] Ibid., p. 23.

[9] See Eric Funkhouser, "Do the Self-Deceived Get What They Want?" *Pacific Philosophical Quarterly* 86 (2005): 295-312, pp. 297-98.

(and presumably other cognitive endorsements) that is not mediated by an agent-style explanation. Second, at least some cases of twisted self-deception are plausibly ones in which there is no such desire.

The most detailed partial story of how desires can have a non-agent-style influence is due to Mele. To begin with, he notes that beliefs may be supported by *positive misinterpretation*, in which evidence counter to the belief is given a good spin (the rejection of an article by a journal is taken to be sign of its substantial challenging character); *negative misinterpretation*, in which counterevidence is reinterpreted to show that something is faulty about the evidence (the referee's report doesn't show the article to be of poor quality but rather displays misunderstanding); *selective focusing or attending* to likely sources of positive evidence, and *selective evidence gathering*.[10]

Mele suggests that selective focusing on, or attending to, likely sources of positive evidence for p can be explained by our desire that p be the case but that the other kinds of support of belief need another type of explanation.[11] Regarding the second point, if there is pleasure to be had in focusing on evidence for p, there is also pleasure to be had in interpreting things in the right way for p to be true and, to an extent, the prospect of pleasure can explain the selection of certain kinds of evidence. Mele seems to have arrived at the impression to the contrary because he compared selective focusing on positive evidence for p with general acts of interpretation and selection. Unfortunately, even though Mele was insufficiently optimistic about the prospects of explanation here, such explanation as we do have seems unlikely to be demonstrably independent of a desire to believe that p. So it cannot play the role of illustrating an alternative kind of explanation to one that is rooted in the latter desire.

More promising materials derive from Mele's discussion of various types of cold biasing of reasoning and what he dubs the "Friedrich-Trope-Liberman (FTL) model" of belief formation.[12] Mele notes that if information is vivid—engaging the imagination—then it has a greater influence on what we believe. Similarly, and presumably relatedly, we take the ready availability of information as a good estimator of the likelihood of events that it concerns. We also tend to search for confirming instances of a hypothesis that we are assessing. Although all these are faults of reasoning that occur cold—that is, without the influence of motivation—it is plausible that if you very much want something to be the case in the world, this will engender exactly the kind of biased process-

[10] Mele, *Self-Deception Unmasked,* chap. 2.

[11] Ibid., p. 28.

[12] Ibid., p. 31.

ing just described. Information concerning what we desire, for example, is much more vivid and readily available.[13] So we have some understanding of how the desire that p may give rise to the belief that p without having to assume that this is mediated by a desire to believe that p.

One caveat regarding the appeal to the mechanisms involved in cold biasing is that they don't, in fact, identify an alternative explanatory scheme in which desire may figure to supplant the appeal to an agency explanation and hence, given the move made above, potential appeal to desires for cognitive endorsements. Although it seems plausible that a desire that p influences the various factors identified without them, we do not have a detailed explanation of how it does. While it is extremely doubtful that an agency explanation should be given for why certain information is vivid to us, we have nothing in its place except brute causality between various mental elements. At best, we just have more details as to how brute causality may operate.

There is another gap in the explanation that is more damaging. Suppose that we desire that p and fear that not-p. It is plausible that this is a classic situation in which self-deception may occur. Fearing that something is the case presumably makes information about whether it is the case more vivid and information about it more readily available. Yet, in many cases of self-deception, what we fear is not believed and what we desire is believed. It appears that appeal to these kinds of factors alone cannot explain what is going on. There is another element.

Appeal to the FTL model of belief formation promises to provide a way around this problem. According to the model, lay hypothesis testers are more concerned to minimize or eliminate "costly" errors than to seek truths. Friedrich calls this the PEDMIN (primary error detection and minimization) analysis of lay hypothesis testing.[14] Friedrich—in considering the application of this approach to self-deception—suggests that hypotheses that lower one's self-esteem (e.g., that one is a fool) are costly. Hence we set the confidence level relating to accepting them high to reduce the chance of believing them falsely.[15] We must have very great evidence that we are fools before we are likely to believe it. By the same token, we set the confidence level for rejecting the hypothesis low so that we increase the chance of rejecting it.

The solution suggested by the FTL model to the problem raised concerning cold-biasing features is that while the desire that p and the fear that not-p render the supporting data of each vivid and readily available,

[13] Alfred R. Mele, *Irrationality: An Essay on* Akrasia *and Self-Deception* (New York: Oxford University Press, 1987), p. 145.

[14] Mele, *Self-Deception Unmasked,* p. 31.

[15] Ibid., p. 34.

the costs associated with falsely believing that p may be higher than the costs associated with falsely believing that not-p. Hence the FTL model would explain why we believe that p rather than not-p.

As before, the FTL model offers only a partial explanatory role for desire. It does not proffer anything other than brute causality to explain the influence of motivation on confidence thresholds. As Mele puts it, motivation "triggers and sustains the operation of a habit that in itself is purely cognitive" and may itself be "automatic and inflexible."[16] However, this is not damaging in itself.

It would also be a mistake to assume that the confidence levels must be set by a desire to cognitively endorse or reject the hypothesis they concern.[17] If other desires than the desire that p are operative, they may be conditional and influence the confidence levels comparatively. For instance, the confidence levels for a belief that p may be set low if the desire to *believe that p if p* is stronger than the desire to *believe that not-p if not-p*. This needn't imply that such a subject desires to believe that p tout court.

More problematic is the fact that this model seems unable to capture the theoretical irrationality of self-deception. The reason for this depends upon whether we take the model to characterize how we ought to form our beliefs or how we do form our beliefs. If it determines how we ought to form our beliefs, then its explanation of self-deceptively formed beliefs entails that they are rational. When a self-deceived subject's beliefs depart from what we suppose might be supported by the evidence, this just demonstrates that they have different confidence levels and, with regard to those confidence levels, form the beliefs they ought to form. It is one thing to allow that self-deception may, on occasion, be beneficial. It is quite another to explain it in such a way that, by the agent's lights, it not only is always beneficial but also reflects the appropriate influence of standards of theoretical reason.

On the other hand, if, by our own lights, we feel we *ought* to be evidentialists regarding belief formation and the FTL model is just a descriptive model about how we do reason, then we are back with a puzzle of a familiar kind. How can self-deceivers consciously depart from evidentialism so that the FTL model correctly describes their strategies of belief formation? The model presumes that there is an answer to this rather than suggesting what it is.

One conclusion that might be drawn from this discussion is that the attempt to identify a non-agency explanatory role for the desire that p, rather than the desire to cognitively endorse p, has proved unsuccessful.

[16] Ibid., p. 32.

[17] See, e.g., Nelkin, "Self-Deception, Motivation, and the Desire to Believe," pp. 394-95.

That is not the lesson that I hope will eventually be drawn. Rather, the discussion in this section reveals that another element is needed in the proper characterization of self-deception. This, I shall argue in section 3, is instability in the face of attentive consciousness. So the first line of resistence to theories that appeal to desires for cognitive endorsements relies upon that discussion. With that, let me turn to the second line of resistence.

The second consideration I noted in favor of appealing to a desire to cognitively endorse p was that it promised to unify various cases of self-deception. More particularly, it is supposed to identify what is common to normal and twisted cases of self-deception in which one self-deceptively cognitively endorses what one does not want to be the case. The classic example of the latter is that of a jealous husband who falsely believes that his wife is unfaithful on the basis of slight evidence showing all the signs that he is very unhappy that this is the case.[18]

Dana Nelkin suggests that the jealous husband desires to believe that his wife is unfaithful because, say, he doesn't wish to be taken to be a fool.[19] The attribution is implausible for a number of reasons. First, not wanting to be taken to be for a fool might mean that he places the confidence threshold for a belief that his wife is unfaithful low, but that is not the same thing as wanting to believe it. He might want not to believe it and hope that the threshold is not passed. Second, if he does not want it to be the case that she is unfaithful it is hard to see why, in this case, he should want to believe that it is so anyway. What is unattractive is also unattractive to believe. This is compatible with not wanting to believe that she is faithful when she is not. However, this latter desire does not imply that he wants to believe that she is not faithful. The different strengths of conditional desires I discussed earlier would serve to characterize the jealous husband's state.

The natural move to make against these criticisms undermines the motivation for adopting the desire for cognitive endorsement approach. The advantage was supposed to be unity, where otherwise we had to recognize that a range of different emotions could have a direct influence upon our cognitive endorsements. The defense of the desire for cognitive endorsement approach rests upon the claim that whenever we have these other emotions, it is always legitimate to attribute a desire in an extended sense to cognitively endorse p. Different emotions are different ways of being pro the truth of p. The problem with this is that it supplies unity in name only. The desire for the cognitive endorsement that p has a number

[18] David Pears, *Motivated Irrationality* (Oxford: Oxford University Press, 1984), pp. 42-44.

[19] Nelkin, "Self-Deception, Motivation, and the Desire to Believe," p. 395.

of different ways in which it may be realized—the various cognitive endorsement-influencing emotions—and there is no commonality except for the fact that the emotions in question influence our cognitive endorsements. Proponents of the defense face a dilemma. Either this commonality is sufficient unity, in which case there is no unity argument against recognizing that different emotions may give rise to beliefs since the unity in question is citable by them too. Or this commonality is insufficient, in which case an appeal to the desire for cognitive endorsement is vitiated.

The situation is different if the various emotions contribute toward an agency explanation by giving rise to an intention to cognitively endorse p. The intention provides the commonality against a backdrop of difference. So agency accounts have the prima facie virtue of covering both ordinary and twisted cases of self-deception.

The general point is not that self-deception never rests upon a desire to cognitively endorse p. If the agency approach is correct for some cases of self-deception, then such an attribution is plausible for them. The point is simply that not all cases should be so characterized. We are left, then, with the following upshot. Appeal to agency explanations appears to be able to capture the commonality but is not mandatory. There seems to be the possibility of non-agency influence though, for the reasons identified, there are still important gaps in the account. Appeal to desires for p alone cannot capture the commonality, and appeal to desire to cognitively endorse p both fails to capture the commonality and is independently implausible. If identifying a common structure is desirable, though, it seems that the agency approach is ahead on points. It unifies more than the other approaches, and while the relevant attributions of states do not seem mandatory, on the other hand they are not ruled out. In the next section, I subject this claim to unity to scrutiny.

2. Agency and Semi-Agency

In the previous section, I noted the apparently more unifying character of agency approaches over those that appeal to the direct operation of the desire that p, or the desire to cognitively endorse p. The burden of this section will be that just as with the appeal to an extended sense of desire mentioned at the end of the previous section, the apparently common appeal to agency masks considerable diversity. Either we recognize that there are many different appropriate characterizations of broadly agent-like activity with regard to the production of a cognitive endorsement that p, or we recognize a further kind of approach involving relationships of semi-agency and the like. Cognitive endorsements produced as a result of the operation of semi-agency will often seem less paradoxical because they imply that agents will be less aware of what they are up to. As remarked in the introduction to this paper, this has its risks. It will be less obvious that we have cases of self-deception. In the subsequent section, we will see how this general concern may be answered by an appeal to instability.

Talk of semi-agency turns on what is required in order to have an intention to cognitively endorse that p—the presence of intention being the hallmark of agency. Suppose that I would like to believe a particular proposition—for instance, that the party was a great success—and I review various pieces of evidence in its favor. I believe that reviewing these pieces of evidence will enable me to conclude that the party was a great success *if anything does*. I have no view about whether the evidence will prove enough. As things turn out, the evidence is sufficiently favorable and I conclude that the party was a great success. My assessment of the evidence is, in fact, skewed by my desire to believe that the party was a great success in some of the ways Mele describes.

The details given are not sufficient for the case to count as a one of self-deception. In addition, the instability identified in the next section is required. Setting this aside for now, the important points are, first, that I did not intend to acquire the belief that the party was a great success yet, second, I did what I did as a means to have the belief that the party was a great success. So there can be cases in which subjects do something as a means to a certain end but do not intend the end or intend to bring about the end.[20] I don't intend to bring it about that I believe that the party was a great success, since what I do is not something that I expect to have the upshot that it does, and so it does not form part of a planned sequence of activity by which I aim to bring about the belief.

The conclusion just reached is independent of the suggestion that a necessary condition of intending to do A is that one select a reliable means by which to do it. Suppose that I throw a dart at the bullseye and it hits. I may have aimed to hit the bullseye and, indeed, tried to marshal my muscles and throwing action into a bullseye-hitting form. But being a very poor (actually, astonishingly poor) darts player, I had little expectation that I would succeed. Did I intend to hit the bullseye? Opinions differ. Some say "no" on the grounds that the means I selected to arrive at the belief were not the exercise of reliable skill.[21] Some say "yes."[22] In

[20] For the distinction, see Alfred R. Mele, "Mental Action: A Case Study," in Lucy O'Brien and Matthew Soteriou (eds.), *Mental Actions and Agency* (Oxford: Oxford University Press, forthcoming).

[21] Alfred R. Mele and Paul K. Moser, "Intentional Action," *Noûs* 28 (1994): 39-68.

[22] E.g., Christopher Peacocke, "Intention and Akrasia," in Bruce Vermazen and Merrill B. Hintikka (eds.), *Essays on Davidson: Actions and Events* (Oxford: Oxford University Press, 1985), pp. 51-73, at p. 69.

the situation described above, the means by which I tried to form a belief may be perfectly reliable. Given my motivational states, looking at the evidence in favor of the party being a great success would generally be sufficient to determine that I believed it was so. Self-deception can be a skill at which we are only too unwittingly adept. The reason why it seems a mistake to take the above case to be a case of intending to produce a belief that p is that we don't have a plan that we are seeking to implement made up of steps leading to the result that p. The final production of the belief is unplanned because I have no view about whether I may succeed in the case in question by doing what I am doing.

In later work, Mele provides considerations that may provide a different verdict regarding the party case. He writes that we may circumvent the issue of whether ascriptions of intentions require reliability by

> focusing on whether people who acquire motivationally biased beliefs that *p try* to bring it about that they acquire beliefs that *p*, or try to make it easier for themselves to acquire these beliefs. If they do try to do this, one need not worry about whether the success of their attempts owes too much to luck or to factors beyond the agents' control, for it to be true that they *intentionally* brought it about that they believed that *p*.[23]

The case I described certainly seems to fit this characterization. For my purposes, it does not much matter whether we tie intentions to reliability and/or planning, or adopt this more liberal suggestion. The important point is that we have a variety of ways in which a subject may be related to the self-deceptively produced cognitive endorsement.

Moreover, there are non-borderline cases in which we seem to do something as a means of producing a belief even though we don't intend to do this. The possibility of nonintentional means-end activity is revealed in other areas. Sometimes when we have an itch or a tickle, we scratch ourselves or move our limbs with the belief that by so doing, the itch or tickle will be alleviated. Here it seems clear that we intend to alleviate the itch or tickle. On other occasions, we still act, but our action is more responsive. It involves no explicit appreciation of why we are doing what we are doing. The itch, as it were, invites and structures our response rather than eliciting a plan of action. In such cases, it still seems to me correct to say our responses are a means to the alleviation of the tickle or the itch. So there is a gap between some action being a means to an end and it involving an intention.

Suppose my motivational state makes a certain hypothesis attractive to me to believe and, hence, I enjoy contemplating it. My implicit belief about what constitutes evidence for that hypothesis results in my enjoyment in contemplating evidence for that hypothesis and I am inclined to

[23] Mele, *Self-Deception Unmasked*, p. 15.

dwell on the evidence because it enables me to feel more and more convinced that the hypothesis is true. I shy away from considering evidence conflicting with the hypothesis because of my implicit expectation that the evidence will be unenjoyable to contemplate. I would say that in such circumstances, I might not intend to produce the belief in the hypothesis, but my activities are a means by which I produce the belief. In these circumstances, means-end beliefs, beliefs about what constitutes evidence for what, and motivational states serve to explain a piece of belief production.

It might be argued that there is an underlying unity. It is just a mistake to seek to characterize this in terms of intentions. There is a weaker notion of agency or semi-agency common to all. However, this is not correct. Sometimes appeal to intentions is crucial and nothing weaker will do.

Suppose that a subject wishes that his paper were finished just as strongly as he wishes, later, that it is *wrongly* rejected. In the first instance, he carries on searching for counterexamples and does not arrive at the belief that it is finished, whereas in the second case he forms the belief that it is wrongly rejected. What explains the difference given the equal motivational strength? William Talbott and José Bermudez have argued that an intention is necessary in at least some cases.[24] The subject did not intend to form a belief as a result of his desire that the paper was finished but did intend to form the belief that the paper was wrongly rejected given his desire that it was wrongly rejected.

Mele responds to this type of case by making two points. First, an explanation can be offered in terms of the different costs in believing that one's paper is wrongly rejected and that one's paper is finished.[25] He justifies this response by noting that there would have to be some difference between the two situations to explain why we formed an intention in one and not in the other. Given that this is so, we can explain why the subject formed the belief that his paper had been rejected unjustly, whereas he failed to form the belief that the paper was finished, just by appealing to this difference. Second, agency views have their own version of the selectivity problem, because there will be occasions when, in the absence of consequent malfunctioning, intentions don't result in what is intended.[26]

I will discuss these points in turn. Regarding the point about different costs, the thing to stress is that we are familiar with situations in which

[24] William J. Talbott, "Intentional Self-Deception in a Single Coherent Self," *Philosophy and Phenomenological Research* 55 (1995): 27-74, pp. 60-63; José Bermudez, "Self-Deception, Intentions and Contradictory Beliefs," *Analysis* 60 (2000): 309-18, pp. 317-18.

[25] Mele, *Self-Deception Unmasked*, p. 63.

[26] Ibid., p. 66.

motivation cannot serve to explain why we have one intention rather than another—for example, Buridan's ass.[27] An explanation of why the ass goes to one pile of oats rather than another exactly similar pile of oats is that the ass intended to go to the first pile of oats. By the very nature of the case, there is no further explanation of the intention that can justify why the first rather than the second pile of oats was sought. It is true that there may be some explanation of why intention to pursue one pile was formed but the explanation would not figure at the rational level. It might just be the result of the firing of a certain batch of neurons, for instance.

Of course, just because intentions have a theoretical utility in explaining why Buridan's ass doesn't starve, it doesn't follow that they need have a theoretical utility in dealing with the selectivity problem. My point is just that it is perfectly legitimate to argue that a resolution of a certain range of selectivity problems will involve appeal to intentions to produce beliefs. Perhaps Mele will deny that there are potential Buridan cases for belief, but it doesn't seem to me that this is right. I can want to believe that I am successful just as much as I want to believe that I am modest. Nevertheless, it may seem that I can't proceed toward one belief without getting further away from the other. So it is, at least, *possible* that an agent's beliefs can display a structure that invites an intentional solution.

Let me now consider Mele's claim that the agency view has its own version of the selectivity problem in which an intention acted upon by one agent is not acted upon by another. Here I just underline a point already made. The simple fact that there might be a selectivity problem that arises for intentions does not mean that appeal to intention is illegitimate to break the tie between motivations and explain why one had a consequence that the other did not. Moreover, Mele's illustrations of cases in which intentions fail to result in actions are ones in which although there is no malfunction of the intention, there is an error in implementation, for example, intending to hit the tennis ball and missing. Recognition of the role of intention in Buridan cases—and its executive role in coordinating action—allows for the possibility that there might be a particular way of resolving the selectivity problem that cannot simply be cashed out in submental terms. So even if it is the case that there are selectivity problems resolved by citing the presence or absence of successful implementation conditions, that does not mean that there are no selectivity problems that should be resolved by appeal to intention.

Mele might argue that in the case of belief, intention has no role to play because arriving at a belief is not an action. Suppose, for the sake of argument that this is true when beliefs are founded on the basis of evi-

[27] Michael E. Bratman, *Intentions, Plans and Practical Reason* (Cambridge, Mass.: Harvard University Press, 1987), pp. 11-12.

dence. Nevertheless, when beliefs become responsive to desires, and in particular competing desires purportedly providing practical reasons for belief, then it is legitimate to suppose that intention may have a role to play. Since Mele is prepared to concede that intention *may* have a role in producing belief, he seems in no position to reject the considerations offered here on the ground that the production of a belief is not an action.

Attribution of intentions in self-deception have a distinct explanatory role contrary to what Mele suggests. Consider a case that Mele discusses. Suppose that Sam has evidence that Sally, his wife, is having an affair and that he favors the belief that she is not having an affair. Sam either does not attend to evidence that she is having an affair or misinterprets it in the ways described above.[28] Mele explains Sam's behavior as follows. First, the confidence level required for believing that Sally is having an affair will be set higher because of the costs to Sam of believing that she is having an affair. Second, he may intend not to focus on the possibility that Sally is having an affair—and hence not dwell on the evidence—because he finds it painful.[29]

Looking at some of the pieces, perhaps we can agree with Mele. If Sam misinterprets the evidence for p, this need not be done in order to believe that p. Similarly, if one intends not to focus upon the evidence, the intention may stop there. Again, one need not intend to believe that she is not having an affair. Nevertheless, there are some aspects of Sam's behavior that suggest the presence of an intention to form the relevant belief. First, there is Sam's avoidance of evidence against the belief, for instance, the avoidance of places where Sally and her lover actually meet. In such circumstances, Sam must, at least implicitly, know that there exists possible evidence that suggests the truth of the proposition that Sally is having an affair and be seeking to avoid this. Sam cannot be responding to the pain of contemplation, because the whole point is that he doesn't feel the pain because he doesn't put himself in a position to feel it. Instead, he must recognize, again at least implicitly, that he might feel pain in such and such circumstances.

Then there is the systematic character of Sam's activities, involving intentionally focusing on and avoiding other bits of information, together with the fact that Sam has beliefs about their evidential merit and likelihood of causing pleasure or pain. Suppose that Sam thinks on a particular occasion "Oh, I don't want to think about that because I might begin to suspect Sally of having an affair." I don't think it would be appropriate to attribute to Sam the intention to produce or sustain the belief that she is not having an affair just because of that. Sam need not, on this occasion,

[28] Mele, *Self-Deception Unmasked*, pp. 57-58.

[29] Ibid., pp. 56-59.

think that the belief was ever in significant danger of being undermined. It is when Sam's behavior takes on a systematic character that it becomes more and more plausible to ascribe the intention. The pattern of information retrieval and avoidance can be characterized as following the guidance of a plan and there is no reason to doubt that it may be the display of a reliable cognitive-affective skill over which the agent has control.[30]

It is a mistake to suppose that the attribution of an intention is discredited because, piece by piece, alternative explanations can be provided of the various elements of Sam's behavior. Proponents of such attributions need not deny that other explanations are possible. Instead, they will emphasize that their explanation is best. William Talbott makes a related point when he talks of the difficulty for Mele's approach in accounting for the resourcefulness of the self-deceived.[31] I think that those who resist this point implicitly rely upon the alleged paradoxes of self-deception to discredit the attributions favored by the agency view. But, as I indicated earlier, if we allow that the allegedly problematic combination of states is, in principle, possible, then there is little support to be derived from such an appeal.

The varieties of agency-like explanation reveal that the apparent unity identified in section 1 is misleading. There is considerable difference. However, all is not lost for the proponents of a unity rooted in agency. They might argue that minimal agency or semi-agency in the production of a belief is essential for self-deception. The other elements are just additional features of particularly sophisticated cases of self-deception. In the next section, I shall argue that this is incorrect and that the proper common feature is a certain kind of instability.

3. Essential Absence of Attentive Consciousness

In brief, my point will be that the question of whether someone is self-deceived turns on the question of whether a certain kind of awareness of the evidence for or etiology of the belief would extinguish it.

What are the grounds for thinking that the question turns on what I say? There is a range of cases in which this feature seems to determine whether we have a case of self-deception. I have discussed these cases in detail elsewhere.[32] I sketch them briefly here.

First, there are those who self-consciously take their belief in God to

[30] See Mele and Moser, "Intentional Action," pp. 57-63.

[31] William J. Talbott, "Does Self-Deception Involve Intentional Biasing?" *Behavioral and Brain Sciences* 20 (1997): 127.

[32] Paul Noordhof, "Self-Deception, Interpretation and Consciousness," *Philosophy and Phenomenological Research* 57 (2003): 75-100.

involve a leap in faith not supported by the evidence. They may also appreciate that their belief in God stems from the way their motivational states influence their belief-forming processes. They are happy to recognize this because fundamentally they approve of their belief in God. It seems to them psychologically and spiritually the right thing to believe. Indeed, the possibility of treating the evidence relating to God's existence as just evidence to be assessed in disinterested terms seems to fail to capture the importance of this belief for them.

Second, there are those parents who don't take themselves to have any special evidence of their son's or daughter's innocence of a crime but they believe it all the same. It seems psychologically and morally the right thing to do: to have faith in their child. Again they may be aware of the role of their motivational states in supporting the belief in question, in part by manipulating how they deal with information relevant to their belief, but once more this seems appropriate. To weigh up the evidence in favor and against their son's or daughter's innocence seems an inappropriately cold thing to do.

Third, there are those people who persist in believing that someone loves them in spite of evidence to the contrary because they keep faith with the ideal of the early character of the relationship. Such individuals may be well aware of how their motivational states affect their beliefs and, indeed, acknowledge that this involves reinterpreting the evidence in favor of the belief that they are no longer loved. Nevertheless, they feel that it is psychologically right for them to persist in believing that they are still loved. Once more, just weighing up the evidence in favor and against the belief that they are still loved seems an inappropriately disengaged viewpoint upon the question of the relationship.[33]

Let us presume for the sake of argument that in all of these cases, the belief produced is false. The inappropriate treatment of evidence is present and it is rooted in desire. It is safe to assume that the body of evidence possessed by the subject favors not the motivationally produced belief, but its opposite. All the conditions for accounts that root self-deception in motivation or semi-agency are met. Yet, none of these is a case of self-deception.

It is no doubt true that the subjects in question are deceived. It is also no doubt true that they are the deceivers and so they are self-deceived in an extended sense. Nevertheless, I deny that they are self-deceived in the important sense that everybody has in mind when they talk about self-deception. My reason for this is that the subjects in question may know precisely how their motivational states affect their beliefs and yet accept it. One way of capturing the point is to say that while they might be de-

[33] Ibid., pp. 76-83.

ceived by themselves they need not be deceived or in ignorance about themselves.

Richard Holton suggests that self-deception always involves an error about oneself. This is the basis of his rejection of Mele's analysis.[34] This seems to be too strong and too weak. It is too strong because, in order to be self-deceived, one need not be mistaken about oneself but merely in ignorance about the particular way in which a cognitive endorsement is produced. The crucial point is that the ignorance or mistake is required in order for the cognitive endorsement to be produced successfully. Holton's position seems too weak because he does not consider the possibility that subjects should know what they are up to but this knowledge is unavailable to consciousness. Nelkin has suggested that lack of awareness might be psychologically necessary for the root belief and desire of a project of self-deception to play their role.[35] I think this is too weak. It is not just psychologically necessary but essential if we are to have a genuine case of self-deception.

Thus I suggest that a necessary condition for self-deception is this:

(a) The subject, S, fails to attend consciously in a certain way, W, to either the evidence that rationally clashes with the, standardly, motivationally favored proposition that she cognitively endorses or some element of the psychological history characteristic of the self-deception behind the cognitive endorsement of the motivationally favored proposition.

(b) If the subject were to attend consciously in way W to both the, standardly, motivationally favored proposition and either the evidence that rationally clashes with it or the psychological history (whichever applied from clause (a)), the, standardly, motivationally favored proposition would no longer be believed.[36]

The condition appeals to attentive consciousness because if, say, the evidence or psychological history just had a phenomenal impact in consciousness without a subject focusing on its nature, then the condition would be much too strong. A self-deceptively cognitively endorsed proposition need not be undermined in such circumstances. Equally, it is important that the subject is attentively conscious of the evidence and/or psychological history appropriately individuated, for example, in terms of propositional contents and kinds of propositional attitudes. Otherwise,

[34] Richard Holton, "What Is the Role of the Self in Self-Deception?" *Proceedings of the Aristotelian Society* 101 (2001): 53-69, pp. 59-60.

[35] Nelkin, "Self-Deception, Motivation, and the Desire to Believe," p. 395.

[36] Noordhof, "Self-Deception, Interpretation and Consciousness," pp. 87-88.

again, the condition would be too strong. Note that I do not have to claim that all attentive consciousness to the evidence or psychological history would undermine the self-deceptively produced cognitive endorsement. It may be that it is possible to attend consciously in a brief or sloppy manner. The necessary condition for self-deception is that there is an absence of attentive consciousness in a certain way that would play this role.[37]

The second clause captures the way in which the ignorance or false belief is relevant to the success of the self-deception. The first clause captures the fact that in self-deception, it is the clashing evidence or motivational support that is not subject to conscious attention rather than the cognitive endorsement produced. If we consider cases similar to the ones I have described in which failure to attend to evidence or some part of the psychological history leading up to the belief is crucial for its retention, then a verdict of self-deception is far more plausible. The reason why the appeal to a weaker notion of agency, or semi-agency, canvased at the end of the previous section does not work is because the question of whether such operations amount to a case of self-deception will depend upon the truth or falsity of the condition just identified.

I have mentioned that the cognitive endorsement that is the product of self-deception has as its content a proposition that, standardly, is motivationally favored. Two kinds of cases suggest that it is not invariably motivationally favored. The first involves twisted cases of self-deception discussed earlier. The second, identified by Martha Knight, appears to involve no motivational states but stems from a particular habitual cognitive schema such as self-underestimation or over-critical holding oneself responsible.[38]

One of Knight's examples is Dolores, who has the idiosyncratic belief that her child died of leukemia because she didn't isolate her cancer-ridden cat from her child. Mele rejects this example as a case of self-deception on the grounds that her impartial cognitive peers—those cognitively like Dolores but without any biasing motivational factors—would arrive at the same belief.[39] This places too much weight on one way in which subjects might be biased. An alternative notion of impartiality excludes the kind of biased reasoning that is Dolores's hallmark by appealing to a moderate idealization of Dolores's cognitive capacities removing the influence of such schemas. If subjects have a systematic propensity to make certain errors in reasoning due to a cognitive schema

[37] I am grateful to Al Mele for raising this possibility.

[38] Martha Knight, "Cognitive and Motivational Bases of Self-Deception: Commentary on Mele's *Irrationality*," *Philosophical Psychology* 1 (1988): 179-88.

[39] Mele, *Self-Deception Unmasked*, pp. 107-8.

that they would control if brought to attentive consciousness, then it is plausible that these will generate cases of self-deception. The fact that subjects would not endorse the outcome of the schema if its operations were brought to attentive consciousness provides the motivation for considering this further idealization.

Mele objects that we would not classify people as self-deceived if they were prone to making simple arithmetical mistakes that they would correct if brought to attentive consciousness.[40] This reveals the importance of the kind of mistake involved in the case of Dolores, or in cases of underestimation such as subjects' cognitive endorsement that they are unattractive and fat (when, in fact, they are at worst of upper-end normal weight). Cognitive schemas of the kind indicated are not going to be behind simple arithmetical mistakes. The failure to attend consciously to elements of the arithmetical problem is not rooted in a cognitive schema or motivation that favors a particular answer at which the subject arrives. That is not to say that subjects cannot make mistakes in arithmetic because they are emotionally distracted, but simply that when they are, the distraction does not favor a particular answer. Moreover, in most cases, it is not clear that failure of attentive consciousness of the indicated kind is required for arithmetical mistakes. Subjects may be perfectly attending to the numbers they need to add up (say) and yet make the mistake.

Proper characterization of the schemas responsible for self-deception is clearly a work in progress. The central cases are likely to involve questions of self-assessment and to have motivational consequences. Nevertheless, this does not imply that they are simply another kind of motivational biasing.

The analysis of self-deception I favor, thus, holds that:

S is self-deceived that p if and only if S cognitively endorses that p and

(a) The subject, S, fails to attend consciously in a certain way, W, to either the evidence that rationally clashes with p, which she believes, or some element of the psychological history characteristic of the self-deception behind the cognitive endorsement that p.

(b) If the subject were to attend consciously in way W to both p and either the evidence that rationally clashes with it or the psychological history (whichever applied from clause (a)), the subject would no longer cognitively endorse p.

(c) (a) holds because of S's motivational state or emotional state that p or cognitive schema that favors the cognitive endorsement that p.

[40] Ibid., pp. 105, 109.

Let me make five comments about this proposal. First, an agent's failure to attend consciously to the cognitive endorsement supporting mental apparatus cannot be explained by appeal to the setting of confidence levels, since confidence levels concern the treatment of evidence and not mechanisms supporting cognitive endorsement. Nor can this lack of attention be explained by motivational bias, since, after all, the mental apparatus is cognitive endorsement-supporting. One would have thought that greater attention would be paid, if anything.

Second, the approach captures the irrationality of self-deception. The crucial point is that the agent is clearly not living up to her own ideals of reasoning otherwise conscious attention to how she arrived at her cognitive endorsement would not undermine it. This point is in need of some qualification if, as I have argued for elsewhere, attentive consciousness actually makes more attractive being disposed to act upon what we take to be true than we might reflectively think is merited.[41] However, in general, it is true. We can take the FTL model to characterize how our beliefs are formed (rather than how they ought to be formed) and note that self-deception occurs when subjects' manipulation of confidence levels is not sustainable if they were attentively conscious of what they were up to.

Third, the analysis does not stipulate that the self-deceived person should secretly believe that not-p or that the evidence is sufficient to believe that not-p. Nor does it require the presence of a guiding intention. The account allows that a particular combination of mental states may, on one occasion, count as self-deception, and on another occasion not. The matter is settled by whether there is the requisite dependency of the favored cognitive endorsement upon lack of conscious attention to evidence or distinctive psychological history.

To illustrate this point further, consider Amélie Rorty's case of Dr. Androvna:

> Dr. Androvna, a cancer specialist, has begun to misdescribe and ignore symptoms of hers that the most junior premedical student would recognize as the unmistakable symptoms of the late stages of a currently incurable form of cancer. She had been neither a particularly private person nor a financial planner, but now she deflects her friends attempts to discuss her condition and though young and by no means affluent, she is drawing up a detailed will. Although she has never been a serious correspondent and reticent about matters of affection, she has taken to writing effusive letters to distant friends and relatives, intimating farewells, and urging them to visit her soon.[42]

[41] See Paul Noordhof, "Believe What You Want," *Proceedings of the Aristotelian Society* 101 (2001): 247-65; and Noordhof, "Self-Deception, Interpretation and Consciousness."

[42] Amélie Oksenberg Rorty, "The Deceptive Self: Liars, Layers, and Lairs," in Brian P. McLaughlin and Amélie Oksenberg Rorty (eds.), *Perspectives on Self-Deception*

Rorty's claim is that, in these circumstances, it is plausible to ascribe to Dr. Androvna the unconscious belief that she has cancer. Mele suggests that we should, instead, ascribe the conscious belief that there is a significant chance that she has cancer.[43]

It is not clear that Dr. Androvna's belief that she does not have cancer can survive for long against a belief that there is a significant chance that she does have cancer. Nevertheless, it is plausible that there may be circumstances in which both beliefs persist for a little while, and Mele can trade on this. The question is whether the attribution of the belief that there is a significant chance that she has cancer can capture the self-deceived character of her behavior. Suppose that earlier in her life she had believed consciously that there was a significant chance that she had a heart condition and she did not respond in the ways outlined above. Instead, she behaved normally and underwent medical examination. It proved to be a false alarm. The description of her earlier behavior just given doesn't seem incompatible with her belief that she has a significant chance of having a heart condition. One way of capturing the difference would be to insist that in the second case, unlike the first, although she believes there is a significant chance that she has a heart condition, she doesn't believe that she has one. If that diagnosis is the only one available, then it has unhappy consequences for Mele's insistence that there need be no secret belief in the opposite of the self-deceptively produced belief. There is another alternative, though. For some reason—perhaps she is in a heightened state of anxiety—the belief that there is a significant chance that she has cancer threatens to result in the belief that she does have cancer, whereas the second does not. To avoid this, Dr. Androvna is not attentively conscious to the mental history that gives rise to the belief that she does not have cancer and/or not attentively conscious to her belief that there is a significant chance that she has cancer. Because she is not, there is no need to assume that she would have arrived at the belief that she does have cancer. Nevertheless, because the lack of attentive consciousness is crucial for this in the cancer case but not the heart condition case, we have a case of self-deception in the former and not the latter.

Fourth, it is common ground that the self-deceived subject should often think that it is the case that p, sincerely avow that p, believe that one believes that p, and so on (where p is the self-deceptively favored proposition). This is part of the functional role of the belief that p. Those who deny that subjects believe that p emphasize the fact that the remain-

(Berkeley: University of California Press, 1988), pp. 11-28, at p. 11; Mele, *Self-Deception Unmasked*, p. 71.

[43] Mele, *Self-Deception Unmasked*, pp. 71-73.

ing part of the functional role of the belief is not realized in cases in which it is reasonable to attribute a belief that not-p. By the same token, though, part of the functional role of the belief that not-p will not be realized either, namely, those parts involving thought and sincere avowal. One way of resolving the matter is to consider which part of the functional role of belief is most important and let that determine which belief, if either, it is correct to attribute. Perhaps there is nothing more to do than say that these partial roles are realized and detail the circumstances in which they are manifested. A second way of resolving the matter is to recognize that functional roles are always relative to circumstances. Not every part of the manifestation of the role is to be expected. In circumstances in which the opposite of a certain belief is also believed, an important part of the role will not be manifested. Nevertheless, this does not mean that the role is not realized, nor that it is inappropriate to attribute both beliefs.

The appeal to instability provides a way through this debate. The need to attribute both the belief that p and the belief that not-p is the most drastic way to capture the instability essential in self-deception. Nevertheless, there are cases in which less will do. Consciously avowing that p will be in tension with believing that not-p if attentive consciousness of one’s belief that not-p would result in one sincerely avowing that not-p. If the grounds for avowing that p are the reasons for believing that p, then a belief about the reasons for not-p could not be present in consciousness with the avowal. The correct attribution will depend upon the details of the case.

By the same token, though, it is a mistake to suppose that sincere avowal, or a higher-order belief, is the sole way in which one might be self-deceived. It is plausible that the basis of one’s sincere avowal, or a higher-order belief that one believes that p, is the reasons for p. When we consider whether we believe that p, we consider the reasons for p. Suppose, for the sake of argument, we in fact do not believe that p (as the proponent of the alternative picture being discussed claims). When we consider the reasons for p, being convinced that we believe that p is being convinced by the reasons. If we are convinced by the reasons, then we come to believe that p. It doesn’t follow from this that we don’t also have the belief that not-p. In being convinced by p, there may be ways in which it doesn’t sink in so that we still believe that not-p. The point is simply that while instability can explain how self-deception need not require that we have both beliefs, at the same time, the means by which we can explain this reveals that sometimes the self-deceived will have both beliefs.

Fifth, and finally, it is often noted that consciousness has something to do with self-deception, although others have located its importance in different places. One is with regard to the product of self-deception. It is suggested that self-deception must involve avowal or a higher-order belief that one believes that p (where p is the self-deceptively favored proposition).

It seems a mistake to take either of these options to identify the central appeal to consciousness. There is no denying that avowal is an essentially conscious state, but there is no particular reason why self-deception must involve such a state. Subjects could be self-deceived because they act as if a proposition is true when (say) deep down they believe that it is not or have evidence that it is not. There is no need to insist that they must avow the self-deceptively favored proposition. Absence of conscious relating to the etiology of the self-deceptively favored belief and relating to the opposite belief, where legitimately ascribed, does not imply that the product must be conscious.

Turning now to higher-order beliefs, it is much less clear that these should be conscious, contrary to what their proponents seem to assume.[44] There is no intimate connection between higher-order belief per se and consciousness. Confusion might arise because of higher-order thought theories of consciousness. According to such theories, a mental state is conscious if it is the object of a higher-order mental state such as belief, where the latter is either appropriately caused by the former or there is a disposition for the latter to be present because of the former. However, such theories do not make the higher-order mental state conscious. It is rather that the higher-order mental state makes the lower-order mental state conscious. In the case of self-deception, appeal to higher-order beliefs is supposed to replace mention of lower-order beliefs. That is, the self-deceived are not supposed to believe that p but simply believe that they believe that p. If there is no object state—the belief that p—then the higher-order state cannot make it conscious.

It is true that certain higher-order thought theories of consciousness suggest, as Eric Funkhouser remarks, that there should be a residual or false consciousness phenomenally similar to consciousness of a mental state when the higher-order mental state is false.[45] However, this suggestion undermines higher-order thought theories of consciousness. Now consciousness becomes an intrinsic—and highly contentious—feature of higher-order beliefs. If the aim of higher-order thought theories of consciousness is to explain the nature of consciousness in terms that do not invoke consciousness, there is no room for intrinsically conscious states of the character described.

[44] E.g., Funkhouser, "Do the Self-Deceived Get What They Want?" p. 306.

[45] Ibid., p. 311 n. 27; David Rosenthal, "Two Concepts of Consciousness," *Philosophical Studies* 49 (1986): 329-59, p. 338.

4. Self-Deception and Delusion

If the identified instability is an essential feature of self-deception, we have an interesting way of distinguishing between self-deception and delusion. The latter is present when subjects adopt different principles of evidential reasoning to support a particular belief, or when they fail to accept the application of standard evidential principles of reasoning to the case in question, so removing the instability. Thus it is observed that the delusory belief persists even if subjects recognize that they would not believe it if it were somebody else's belief. For example, a husband with Capgras's Delusion (in which a subject supposes that a loved one is replaced by an imposter) may accept that he would take somebody else's assertion that his wife had been replaced by an imposter as absurd and implausible, and recognize that others will take his belief in similar fashion.[46] Yet such subjects will retain the deluded belief without apparently feeling under any pressure. Funkhouser also remarks on the lack of pressure that the deluded feel but puts this down to the fact that they no longer have the motivationally unfavored belief.[47] I have suggested that the self-deceived don't need to have this belief in order to feel under some degree of pressure. They might believe that the evidence favors the motivationally unfavored proposition. Yet, deluded subjects in such circumstances are unmoved for the reasons mentioned above.

The subjects I described at the beginning of the previous section—the believer in God, trusting parents, and faithful lover—share similarities with deluded subjects insofar as they similarly feel an absence of pressure when faced with counterevidence. The difference between such subjects and the deluded turns on the extent to which their intransigence in the face of the counterevidence stems from principles concerning what we ought to believe that we can reflectively endorse, and the extent to which the subjects are intransigent.[48]

It is instructive to compare this proposal with Mele's account of the difference between self-deception and delusion. In his discussion of Capgras's Delusion, Mele notes that it is plausible that a pair of factors seem to be involved. There are unusual experiences of the loved one, perhaps due to a deficit in the subjects' affective responses connected to perception. They interpret their lack of a reaction as an indication that something is wrong with the person experienced. This component is drawn from Brendan Maher's account of the basis of delusions supple-

[46] See, e.g., Davies et al., "Monothematic Delusions," pp. 149-50.

[47] Funkhouser, "Do the Self-Deceived Get What They Want?" p. 303.

[48] See Noordhof, "Self-Deception, Interpretation and Consciousness," for some preliminary discussion.

mented by some work of Hadyn Ellis and Andrew Young.[49] In addition, there is a cognitive deficit, namely, that such subjects seem unable to adopt a critical stance to their experiences and seriously consider the possibility that they might be misleading. In recognizing the importance of this second factor, Mele follows Martin Davies, Max Coltheart, Robyn Langdon, and Nora Breen while sharing their misgivings about the implications of this idea for the delusory subjects' response to illusions more generally.[50]

Assuming that appeal to something like these two factors is correct, the key difference from self-deception, as far as Mele is concerned, is that delusions are not based in *motivational* factors, whereas self-deception is.[51] Although it may, in fact, be the case that motivational factors are importantly absent in the case of delusions, it is not obvious that this goes to the heart of the matter. The believer in God, the trusting parents, and the faithful lover all have motivational factors present, but because this was accompanied by stability, they seemed not to be cases of self-deception. This suggests that the question of whether there is the appropriate kind of motivational biasing does not settle the question of whether self-deception is present. Indeed, Mele's approach forces him to suggest that cases of delusional jealousy may turn out to be cases of self-deception and that plausible cases of self-deception resting on habitual cognitive schemas alone must be cases of delusion.[52] The case of Dolores we discussed earlier would be a case in point.

My proposal can draw upon the possibility that there are unusual experiences. Their uncritical acceptance as veridical is one way in which subjects may adopt different principles of evidential reasoning or fail to accept that particular application of standard principles. The difference between Mele's approach and my own concerns our different verdicts regarding the centrality of motivational factors and instability. My appeal

[49] E.g., Brendan Maher, "Delusional Thinking and Perceptual Disorder," *Journal of Individual Psychology* 30 (1974): 98-113; Hadyn D. Ellis, Andrew W. Young, Angela H. Quayle, and Karel W. de Pauw, "Reduced Autonomic Responses to Faces in Capgras Delusion," *Proceedings of the Royal Society: Biological Sciences* B264 (1997): 1085-92.

[50] Davies et al., "Monothematic Delusions," p. 153.

[51] E.g., Alfred R. Mele, "Delusional Confabulations and Self-Deception," in William Hirstein (ed.), *Confabulation: Views from Neuoroscience, Psychiatry, Psychology and Philosophy* (Oxford: Oxford University Press, forthcoming); and "Self-Deception and Delusions," in Tim Bayne and Jordi Fernandez (eds.), *Delusion and Self-Deception: Affective and Motivational Influences on Belief Formation* (New York: Psychology Press, forthcoming).

[52] Alfred R. Mele, "Self-Deception and Three Psychiatric Delusions: Robert Audi's Transition from Self-Deception to Delusion: Reflections," in Mark Timmons, John Greco, and Alfred R. Mele (eds.), *Rationality and the Good* (Oxford: Oxford University Press, 2007), pp. 163-75.

to the importance of instability rests on the intuition that the deluded lose grip on reality in a way that the self-deceived do not. For the latter, reality always threatens to break in.

It may be that motivationally supported cognitive endorsements are generally unstable in the way that other kinds of breakdown are not. So there may be agreement about many cases. My suggestion is simply that the essential feature of self-deception, rather than full-blown delusion, is best characterized in terms of the instability. It also provides a natural way of explaining how self-deception can turn into full-blown delusion, namely when instability is banished because the subject no longer endorses or accepts the application of the evidential principles that, in conjunction with their psychological history, placed the favored cognitive endorsement under threat.

5. Concluding Remarks

Self-deception comes in many varieties but at the heart of this variety is an instability that differentiates it from delusion. The instability rests upon two things: first, a particular kind of consciousness that I have characterized as attentive consciousness, and second, the application of principles regarding cognitive endorsement in the light of prior combinations of states (or the propositions that characterize their content). A deeper understanding of the nature of self-deception—and its role in our mental life—will depend upon developing our understanding of these two features. Both are works in progress and promise to illuminate the nature of consciousness and reasoning as much as the phenomenon of self-deception that depends so much upon their absence.[53]

Paul Noordhof
Department of Philosophy
University of York
pjpn500@york.ac.uk

[53] I would like to thank Al Mele for his comments upon a previous draft of this paper.