Do the Self-Deceived Get What They Want?

Eric Funkhouser

Abstract: Two of the most basic questions regarding self-deception remain unsettled: What do self-deceivers want? What do self-deceivers get? I argue that self-deceivers are motivated by a desire to believe. However, in significant contrast with Alfred Mele's account of self-deception, I argue that self-deceivers do not satisfy this desire. Instead, the end-state of self-deception is a false higher-order belief. This shows all self-deception to be a failure of self-knowledge.

Do the Self-Deceived Get What They Want?

Eric Funkhouser

In the last few decades several anthologies and numerous philosophical articles have been dedicated to analyzing self-deception. Though almost all parties agree that self-deception is some type of motivated irrationality, the most fundamental questions about its nature remain unsettled. I aim to go some way towards resolving two of the most basic of such questions: What do self-deceivers want? And what do self-deceivers get? The answers I will offer are in significant contrast with the recently influential account of self-deception offered by Alfred Mele.[1]

Section 1 addresses the question of what self-deceivers want. In the current debate over the nature of self-deception, a division exists between those who judge self-deception to result from an intention to deceive (intentionalism) and those who hold that the deception merely be the result of some motivational state, typically a desire (motivationalism).[2] Though I favor motivationalism, I will not argue against the intentionalist here—motivationalism will simply be assumed. Instead, I am concerned with an internal dispute amongst motivationalists that results from the following question: What is the content of the operative motivational state of self-deceivers?

Two answers have been prominent. Let '*p*' stand for the proposition that the deception is about. These two candidates for the operative desire in all cases of self-deception are:

World-focused desire: Self-deceivers desire that the *world* be such that *p*.

Self-focused desire: Self-deceivers desire that *they* be such that they believe that *p*.[3]

The self-focused version is endorsed in section 1. Those who endorse the self-focused version typically assume, or argue, that self-deceivers satisfy this operative desire by acquiring the belief that *p*. However, in section 2 I argue that self-deceivers do *not* satisfy this desire. An important distinction is made between cases in which this desire is satisfied and cases in which the desire is not satisfied. This is the distinction between self-delusion and self-

deception. (This special terminology is explained later.) While Mele has offered a strong account of self-delusion, his proposal does not fit with the realities of self-deception.

If they do not satisfy their operative desire, what is it that self-deceivers achieve that makes them self-deceived? In section 3 I argue that successful self-deception results in a false higher-order belief that the self-focused desire is satisfied. Self-deception is then a failure of self-knowledge. This account fits the practical irrationality of self-deceivers, explains the cognitive and behavioral "tension" characteristic of self-deceivers, and explains why self-deceivers cannot be fully aware of their self-deception.

1. What self-deceivers want

Given the assumption that self-deception is motivated, we can inquire about the content of this motivation. Considering cases can help us in this endeavor.

> Case 1: Mitchell is a vain man, sensitive about his receding hairline. He has taken to combing his hair over from one side in a rather exaggerated and distasteful manner. Though he takes such obvious steps to disguise his baldness, he fails to acknowledge that he is bald. His friends find it embarrassing that Mitchell often makes a point of mockingly referring to the baldness of other men, while failing to recognize that he is one of them. Mitchell's directions to his barber are very precise, and he invariably poses at a certain angle whenever being photographed or viewing himself in the mirror. Mitchell does not even allow his wife to tussle his hair.

Mitchell, I take it, is self-deceived about his hair. One might think that the motive for this deception is obvious—Mitchell desires that he have a full head of hair. This suggests a general account of the motivation of self-deception. Whenever a person is self-deceived about *p*, that person's self-deception is motivated by a desire that *p*. This is the world-focused version.

As sensible as that suggestion might sound for Case 1, it does not generalize to other cases.

> Case 2: Joey is a jealous man. Objective observers are convinced that it is highly unlikely that his wife Marcia is having an affair. But Joey says otherwise, often calling her names that shock his friends. They tell him that he has no reason to think that she's sneaking off to see another guy whenever she visits with her female

> friends. They say that her recently increased sex drive is probably not an act to cover her guilt. But Joey violently protests. Unsurprisingly, Joey does not like to be proven wrong and refrains from calling Marcia's friends to confirm her stories.

Joey, I take it, is self-deceived about his wife's fidelity. But the motive for his deception is not so obvious. Perhaps some men have unconscious desires to have their relationships sabotaged, but Joey does not—he genuinely wants Marcia to be faithful. The suggested generalization from Case 1 fails because Joey lacks a desire corresponding to the content of his deception. Mele (2001) termed cases of this form *twisted* self-deception. Case 1, in contrast, is *straight* self-deception because the agent desires that what he is deceived about is the case.

The distinction between straight and twisted self-deception suggests two possibilities about motivation. First possibility: Corresponding to these two types of self-deception, there are two versions of motivational content. Second possibility: There is still a unified account of motivation, but it is more sophisticated than we originally thought. Unified explanations generally being preferred, it would be better if we can vindicate the second possibility.[4]

But what motivation is common to both Case 1 and Case 2, as well as all other paradigmatic examples of self-deception? Nelkin (2002) has been the most recent advocate of what I take to be the correct answer. Whenever a person is self-deceived about *p*, that person's self-deception is motivated by a desire *to believe* that *p*. The first step to confirming this proposal is establishing that such desires are always true of self-deceivers. Establishing this for Cases 1 and 2 is a good start.

Mitchell not only desires a full head of hair, he desires to believe that he has a full head of hair (*even if* he does not have a full head of hair). What evidence is there for this latter desire? His avoidance behavior[5] (e.g., always looking in the mirror at a certain angle and his combover technique) suggests that he is not motivated to believe the truth. Rather, he wants to believe that he has a full head of hair because that belief is valuable for its own sake. This

belief is comforting to his vanity, and he wishes to believe it regardless of its truth-value. Joey, too, engages in avoidance behavior (e.g., not checking on his wife's stories) that reveals his desire to believe for non-truth-conducive reasons. But for Joey the deception regarding his wife's infidelity is not intrinsically pleasant. Still, he is motivated to acquire this belief even though the evidence available to him does not support it. Being the jealous, insecure type, and having been cheated on before, Joey desires, *out of caution*, to believe that his wife his unfaithful.[6]

Avoidance behavior and other indications that self-deceivers know the truth "deep down" provide support for the self-focused account of motivation. The best explanation for such phenomena is that the self-deceiver is guided by a desire for a certain state of mind. Perhaps the desire can be given greater specificity in certain cases, but the most appropriate generalization we can truthfully make of self-deceivers, limiting ourselves to the traditional categories of folk psychology, is that they are guided by a desire to believe. But this "guiding" need not be construed as either conscious or intentional. In fact, for most cases it is neither.

Further support for the desire to believe account might be sought by making comparisons to interpersonal cases of deception. When person A deceives some other person B about *p*, A desires that B believe that *p*. Person A certainly need not desire that *p*. The self-focused account preserves this form of motivation even in cases when A=B. However, making this parallel offers little support. Well-known paradoxes arise when self-deception is viewed on the model of interpersonal deception,[7] and theorists must deny a parallel at some point. In section 2 I argue for placing this break elsewhere. But it is not merely accidental that the motivation aligns for interpersonal and self-deceivers. It is the nature of deception to aim at changing someone's mind, not the world.

The desire-to-believe account of motivation suggests a third type of self-deception, in addition to the now standard straight and twisted varieties. In straight cases of self-deception the agent desires that *p*, and in twisted cases the agent desires that not-*p*. But it is

not necessary that self-deceivers have any such world-focused motivation. It is possible for there to be self-deceivers who neither desire that *p* nor desire that not-*p*. That is, one can desire to believe that *p*, while lacking any world-focused desire regarding *p*. Such self-deception would be appropriately termed *indifferent* or *apathetic* self-deception. It should be stressed that such self-deceivers still have some motivation (i.e., a self-focused motivation); they are simply indifferent or apathetic at the world-level. We should expect such cases to be unusual, however, because it would be odd to have a vested interest in believing that *p* without having a vested interest in *p*. But there are such cases. One example is self-deception prompted by peer pressure. We often have desires (more generally: motivations) to be like those around us. Common examples include desires to dress and talk like our peers. Sometimes we even have desires to *believe* as those around us believe. Just as it can be awkward to be the oddball in dress or speech, it is sometimes awkward to hold a minority belief. And one could be motivated to self-deception by having a desire to believe what one's peers believe, while being indifferent to the truth or falsity of what is believed. The belief is simply desired for its utility. Such a case does not fall under either the straight or twisted varieties, but is accounted for by our self-focused desire.

Nelkin (2002) argues that the desire to believe account reveals self-deception to be an understandable example of practical rationality.[8] Self-deceivers have a desire to believe that *p*, and they come to believe it. Self-deception is a species of goal or desire satisfaction, she reasons. In this regard, I think Nelkin's desire-to-believe account is flawed. In the next section we will argue that self-deceivers, or an interesting class of self-deceivers at least, do not get what they want. This case is heavily supported by the presence of avoidance behavior, and other such indicators, we have previously discussed.

2. Do self-deceivers get what they want?

Accurately characterizing the belief states of self-deceivers is no easy task. Mele and Nelkin, in their motivationalist analyses, take it as uncontroversial that self-deceivers are successful in coming to believe as they desire.[9] In our examples, they would interpret

Mitchell as believing that he has a full head of hair and Joey as believing that his wife is having an affair. (Better: Either these interpretations hold, or Mitchell and Joey are not self-deceived.) While Mele spends considerable time arguing against "dual-belief" views[10]—views that interpret self-deceivers as both believing that *p* and believing that not-*p*—he has surprisingly little to say in support of his positive position that self-deceivers believe that *p*. An analogy to interpersonal cases would support attributing this belief to the self-deceived, but Mele certainly is not sympathetic to such analogies.

A number of critics have objected to Mele by noting that there is a tension in self-deceivers that Mele's analysis fails to capture.[11] Robert Audi has characterized the tension of self-deceivers as a tendency to say one thing while believing another.[12] Applying this to Case 1: Mitchell will sincerely assert that he is not bald, though he knows that he is bald. Kent Bach has argued that, even if the self-deceiver does not believe the truth, he at least has nagging suspicions.

> In self-deception, unlike blindness or denial, the truth is dangerously close at hand. His would not be a case of self-deception if it hardly ever occurred to him that his wife might be playing around and if he did not appreciate the weight of the evidence, at least to some extent.[13]

Others have noted that suspicions, coupled with avoidance behavior, are characteristic of self-deceivers.[14]

I agree with Audi that self-deceivers tend to say one thing, while believing another. It would be helpful to have this judgment backed by an account of belief that explains why, in our belief attributions, non-linguistic behavior should be privileged over linguistic behavior (at least in the case of self-deceivers). In particular, why should Mitchell's avoidance behavior count as sufficient evidence that he believes he is bald, but his avowals not count as sufficient evidence for the contradictory belief? Fortunately, we do not need to develop a full theory of belief to answer this question. Two very general and widely accepted claims about beliefs and desires will help us towards this end. The first claim is that beliefs and

desires earn their keep by the role they occupy in predicting/explaining behavior, interacting with other mental states (including other beliefs and desires), and being caused by external stimuli. This is a statement of the Functionalist's position. We will give special attention to the role of beliefs and desires in explaining and predicting behavior—a role that reaches further back, brought to the forefront by the Behaviorists. I assume that there is some necessary connection between beliefs/desires and behavior. Our second claim about beliefs is that the psychological explanation of human behavior is not simplistic. Even assuming a belief-desire psychology, it is not the case that a single belief-desire pair explains any actual human behavior. Rather, it is an entire network of beliefs and desires that interacts to prompt behavior. Because of these complex interactions, it is possible for an agent to have a single belief-desire pair that suggests a certain behavior, but yet the behavior be thwarted by pressures elsewhere in the network.

Given this complexity, it is not possible to look at one tidbit of behavior and infer the beliefs and desires of the agent. (This is the general point of evidence underdetermining theory, as made famous in Quine's applications to the translation of languages and psychological states.) When someone claims "*p*", we generally take this as some evidence that that person believes that *p*. This is because we generally attribute to people a desire to tell the truth *and* assume that they have some privileged access regarding what they believe. So, given the proper motive and ability, the utterances of others can provide us with windows to their psychology. But if either this motive or ability is lacking, then we have a defeating condition. I suggest that when the motive is lacking, they are deceiving us; but when the ability is lacking, they are often deceiving themselves. One way in which this privileged access can become tainted is by the presence of a desire to believe. We can often "find" what we want to find, even if it is supposed to be in our own head. Such a result is to be expected given the well-known confirmation biases and (at least possible) opacity of the mind.

We are now in a position to identify the reason for giving greater weight to Mitchell's non-linguistic behavior. There are other desires in the network that explain why Mitchell's general desire to tell the truth does not combine with his belief that he is bald and cause him to utter, "I am bald!" Namely, Mitchell desires to believe that he is not bald, and this desire explains why he gives utterances supporting the content of that desire. A natural question arises, however: Why doesn't the desire to believe that he is not bald similarly prompt Mitchell to the non-linguistic behavior of one who actually believes this? For example, why doesn't this desire cause Mitchell to quit the combover tactic? An answer to this question should point to a difference between the non-linguistic and linguistic behavior with regard to the desire to believe. The most important difference seems to be as follows. By asserting "I am not bald" Mitchell is not jeopardizing his goal of so believing. If anything, such repetition would further that goal. However, by allowing his wife to tussle his hair Mitchell is jeopardizing this goal. It would be quite difficult for Mitchell to refuse to acknowledge his baldness were that to happen.

This explanation also reveals that Mitchell does believe that he is bald. It is no accident that Mitchell systematically avoids evidence of his baldness. There is no other plausible psychological explanation for Mitchell's avoidance behavior but to attribute this belief to him.[15] However, we have seen that there is another, plausible psychological explanation for his contrary avowals. We know, as a matter of empirical fact, that such desires can cause one to avow as a true believer. Plus, such avowals do not frustrate the desire to believe.

We have seen that we must look at the entire network of beliefs and desires, and consider alternative explanations, when attributing beliefs and desires. Indeed, it sometimes happens in our belief attributions that, contrary to the situation with self-deceivers, avowals that "*p*" will trump behavior generally indicative of a belief that not-*p*. But, again, there will have to be a special story as to why this behavior does not warrant attributing the belief that not-*p*. Here is one such example: I am playing a game of Old Maid with a child. It is obvious to me which card is the "Old Maid", but I pick that card anyway and lose the game.

An adolescent observer later asks how I could fail to notice which card was the "Old Maid." I respond, "Of course I knew. I was just allowing the child to win." When we understand this supplemental desire to allow the child to win, we can then understand why I behaved contrary to how one playing the game generally would (i.e., in a competitive mode) with my beliefs.

It is possible to complete the story about Mitchell in such a way that his combover technique and refusal to let his wife tussle his hair are not sufficient evidence for attributing to Mitchell the belief that he is bald. But, such a tale would require attributing fanciful beliefs and/or desires elsewhere in Mitchell's network of beliefs. For example, Mitchell refuses to let his wife tussle his hair because he believes it is an unmanly thing to allow. And, he opts for the combover technique because he genuinely likes that style. If all of Mitchell's "avoidance" behavior can be explained in such a manner, then we are no longer justified in attributing to Mitchell the true belief. And, as a matter of fact, many self-deceivers eventually do acquire these auxiliary beliefs and desires, and engage in their "avoidance" behavior for these reasons. When this happens, however, they have passed from being self-deceived to self-deluded (a distinction to be explained shortly). For, we are no longer entitled to attribute to them the true belief.[16] As a matter of psychological epistemology, it might not be an easy task to determine when such a transition occurs. But this is because there is a general problem with attributing beliefs and desires, and determining an agent's *real* reasons for action. This is not a special problem for self-deception.

Cases 1 and 2 display the tension characteristic of self-deception—avoidance behavior that points to the agent possessing the true belief, but the agent avowing otherwise. But let's consider a final pair of cases to illustrate the distinction between the presence and absence of such tension.

> Case 3: Nicole possesses much evidence that her husband Tony is having an affair with her friend Rachel. Nicole's other friends have reported to her that Tony's car is often seen parked in Rachel's driveway, at times when he claims to be with his male friends. Tony has lost sexual interest in Nicole, and other suspicious behavior

> provides sufficient evidence for Nicole to be more than skeptical. Yet she laughs off the concerns of her girlfriends, and thinks to herself that Tony is certainly a faithful husband. ("After all, I am still an intelligent, charming, and attractive woman—certainly more so than Rachel!") Yet, in the evenings when Tony claims to be with his male friends, Nicole avoids driving by Rachel's house—*even when it requires her to drive out of her way*.

Nicole's avoidance behavior might be unconscious or it might be rationalized away, but this behavior certainly suggests that she has beliefs contrary to what she tells herself and her girlfriends. Why else would she go out of her way to avoid Rachel's house? Similarly, why else would Mitchell refuse to let his wife tussle his hair? Why else would Joey refrain from confirming his wife's stories?

The tension dissipates if we change the story at the end:

> Case 4: Nicole possesses the same evidence as in Case 3, and she laughs off the allegations just as before. The key difference is that she exhibits no behavior indicating that she suspects, or outright believes, that her husband is having an affair with Rachel. As such, Nicole confidently drives by Rachel's house even when Tony claims to be with his male friends. Why wouldn't she? She believes that he isn't there.

The absence of suspicion and avoidance behavior separates Case 4 from the previous three. Her behavior and accompanying first-person phenomenology whole-heartedly indicate that she believes as she desires.

We can grant that Mele's model, describing the self-deceived as believing as they desire, accounts for cases of the fourth type. But we should question whether cases of the fourth type are what raise the interesting problems of self-deception. Following Mele, we can distinguish between the dynamic puzzle of how an agent can *set about* deceiving himself and the static puzzle of how he can *simultaneously* possess the requisite doxastic states for self-deception. We can agree with Mele's proposal that no intention to deceive is necessary, and that standard biases can resolve the dynamic puzzle. However, there is also a static puzzle regarding how to characterize the belief-states of self-deceivers. While Mele is correct that no such static puzzle arises for cases of type 4, his proposal is inadequate for handling

examples like our Cases 1-3. I suggest that cases of this latter type have provided the philosophically interesting cases of self-deception all along.

The current debate over self-deception would benefit from terminology that can separate these two types of cases. I suggest the following terminology. Let us reserve the term *self-deception* for cases like 1-3. In cases of self-deception the agent has a desire to believe that *p*, and this motivates her to engage in biased reasoning, avoidance behavior, and similar deceptive measures that have been extremely well characterized by Mele and other theorists. For self-deceivers, this desire does *not* result in a belief that *p*, however. (What does it result in? This question is answered in the next section.) Self-deceivers engage in behavior, which reveals that they know, or at least believe, the truth (not-*p*). I hope the cases of Mitchell, Joey, and our original Nicole, have adequately illustrated this combination.

With regard to the static puzzle, Case 4 is uninteresting. Let us call such cases *self-delusion*. Self-delusion is the state self-deceivers enter once they believe what they want to believe.[17] Lest anyone protest, such cases surely are deception in some sense, but we should mark this kind of deception off from our previous category. 'Delusion' seems to capture the full-blown misjudgment which separates the Nicole of Case 4 from the Nicole of Case 3.

This distinction is important to note. When Mele presents examples of (what he calls) self-deception, he invariably presents cases of what we call self-delusion.[18] Such examples include a survey which shows that 94% of university professors think they are better at their jobs than their colleagues.[19] A good many of these professors are presumably guilty of motivated irrationality. Mele's examples of "garden-variety" self-deception are almost all of the same form: parents believe against the evidence that their child has not committed treason or has not experimented with drugs, a scholar inappropriately concludes that his paper should not have been rejected for publication, or a young man misinterprets a woman's behavior as evidence that she is in love with him.[20] Mele is right that in such cases, as he describes them, it is clear that the agent believes what is false. These are cases of self-delusion, and Mele presents a plausible account of self-delusion. But who would have

thought that such cases pose a static puzzle? They clearly lack the necessary ingredient found in our Cases 1-3 that separates self-deception from self-delusion—the presence of behavior that points against the avowed belief.

Mele has considered similar criticisms before, and responds with comments like the following:

> For example, it is alleged that although self-deceivers like Sam [from an earlier example] sincerely assure their friends that their spouses are faithful, they normally treat their spouses in ways that manifest distrust. This is an empirical matter on which I cannot pronounce.[21]

But this response is not sufficient to handle the present objection. My claim is not an empirical claim (or a claim of fiction-interpretation) about whether the characters in Mele's examples really would behave in ways that manifest distrust. They very well might not. But then those would be cases of self-delusion, and why would we think that such cases pose a static puzzle? Instead of such an empirical objection, my point is that there *are* cases in which people sincerely avow one thing while behaving otherwise. These are the philosophically interesting cases that pose the static puzzle. So I suggest that we limit the term 'self-deception' to these interesting cases and reserve 'self-delusion' for Mele-style examples. And Mele does not offer an account of self-deception so defined.

There is no parallel to self-deception, so defined, in cases of interpersonal deception. All cases of interpersonal deception are analogous to self-delusion. In self-delusion the agent believes as he desires, and in interpersonal deception the victim believes as the deceiver desires. Once the self has been deluded, or the interpersonal victim deceived, the deceiver can "walk away" from the act and the deluded state can still remain. Self-deception, in contrast, is a continual process of believing truly, but hiding this belief from oneself out of a desire to believe otherwise.[22]

3. What self-deceivers get

Self-deceivers do not believe as they desire to believe. But there must be some state that they enter into which separates them from those who have a motive to believe but are not self-deceived. What is this end-state that marks the transition from self-deceiv*ing* to self-deceiv*ed*? I suggest the following: self-deceivers (falsely) believe that they believe as they desire.[23]

We find initial support for this suggestion by observing that it seems to capture the irrationality of self-deceivers. Self-deceivers do not get what they want. This type of irrationality is recognized by those who advocate the world-focused version of the motivating desire. Recall that according to the world-focused version the self-deceived desire that *p*, and this then motivates them to acquire the belief that *p*. But this end-state is not a satisfaction of the operative desire—believing that *p* typically does not make it so. On the world-focused version the self-deceived mistakenly believe that the end they desire has been achieved—*p*. (Clarification: The self-deceived need not believe that they have satisfied their operative desire as such. They simply believe that *p*, and this, as a matter of fact, is the satisfying end-state.) Call this type of irrationality *Mistaken Ends Irrationality*.

In section 1 we argued that the world-focused account of motivation is mistaken. Still, such theorists seem to correctly capture the type of irrationality displayed by self-deceivers. Let us now apply Mistaken Ends Irrationality to the self-focused version. On our account the self-deceived desire to believe that *p*. In their self-deception they mistakenly believe that they have attained the end they desire. That is, the self-deceived (falsely) believe that they believe that *p*.

While Mistaken Ends Irrationality does ring true of self-deceivers, the reasoning here might seem a bit too clever. And, we have already rejected the world-focused version of motivation, so why assume that the world-focused account is right in this regard? To better support our chosen end-state, we should explain what is characteristic of higher-order belief failures and how these characteristics are found in the self-deceived.

It takes a reflective creature to have second-order beliefs—a creature that is thinking about its own mental states—and as such there is good reason to associate second-order beliefs with conscious reasoning and introspection.[24] Another reason why second-order beliefs might be more associated with consciousness and cognition, as compared to their first-order counterparts, is that it is difficult to see how a person's behavior could directly reflect the second-order belief that *p* (i.e., the belief that they believe that *p*) apart from simply reflecting the first-order belief that *p*. Indeed, some philosophers have gone so far as to analyze *all* conscious experience in terms of higher-order thoughts or beliefs.[25] While I do not accept a higher-order belief account of consciousness in general, there are certainly significant connections and correlations between higher-order beliefs and conscious awareness. As has been recently argued, higher-order theories are most appropriate for explaining self-consciousness and mental states that we are *aware of*.[26]

I propose that second-order beliefs endow their possessor with four distinct abilities. An agent possessing a second-order belief can: a) report, b) entertain as true, c) use in practical reasoning, and d) integrate in theoretical reasoning the embedded first-order belief. These are not necessarily abilities one possesses simply in virtue of believing that *p*. For example, assume Jill believes that a childhood friend of hers lived in a certain house. Jill might retain this belief from childhood, but only in the remote recesses of her mind. Jill cannot report this belief to anyone, nor does she ever entertain it as true. The belief is not given any consideration. It is an isolated tidbit, removed from the premises Jill uses (even implicitly) in her practical and theoretical reasoning. In short, she does not believe that she believes that a childhood friend of hers lived in a certain house. Returning to her old neighborhood might stir this hibernating belief, causing her to gain the higher-order knowledge. *Then* the first-order belief can be entertained and utilized.

It sometimes happens that people have false higher-order beliefs. Such cases are failures of self-knowledge—having mistaken beliefs about what one believes. I have suggested that this is the case for the self-deceived. The self-deceived do not believe that *p*, but they believe

that they believe that *p*. We might wonder if the four abilities that typically come with higher-order beliefs hold even when the embedded first-order belief is lacking.[27] The first two abilities *do* hold even when the higher-order belief is false. The self-deceived can, and do, report that they possess the false first-order belief and entertain this thought as true. These two abilities are public and private versions of truth-avowals. Mitchell would claim to others, and to himself, that he is not bald, even though he truly believes that he is bald. These first two abilities reside more in the realm of contemplation than action, and thus relate more to the first-person character of beliefs. When one reports and entertains something as true one can give the superficial appearance of genuine first-order belief, as well as possess its characteristic phenomenology. And this is at least part of what self-deceivers want. Their Mistaken Ends Irrationality is such that they desire a certain peace of mind, even if "deep down" they know that the world is not as they desire to believe that it is.

The comments at the end of the last paragraph suggest an alternative account of the motivational content of self-deceivers. Maybe the self-deceived do not desire to believe that *p*, but desire simply to have the first-person qualities associated with believing that *p*. This is roughly the desire to have the second-order belief that *p*. Since this is the end-state that the self-deceived get, on this alternative the self-deceived are not guilty of Mistaken Ends Irrationality. To settle this question we would need to know if Mitchell, Joey, Nicole, and other self-deceivers desire the behavioral dispositions associated with the deceived belief, in addition to desiring the phenomenology associated with the deceived belief. We know that they do not get these behavioral dispositions in any robust sense (e.g., Mitchell won't let his wife tussle his hair and Joey won't check to confirm if is wife is telling the truth). But do both straight and twisted self-deceivers desire such dispositions? Given what we have already established, this is equivalent to asking if self-deceivers are guilty of Mistaken Ends Irrationality. At this point I am willing to hedge a bit and leave this as an open question. (The title is then disappointing if the reader was expecting a definitive answer.) But we have supported some definite answers in the vicinity. The self-deceived desire either to believe

that *p* or merely to have the first-person qualities associated with such a belief. The self-deceived get these first-person qualities, but do not get the belief itself.[28]

Let us return to our discussion of the 4 abilities associated with higher-order beliefs. When the first-order belief is lacking, the last two abilities that typically come with higher-order beliefs—use in practical reasoning and integration in theoretical reasoning—are severely tempered. Nicole, the self-deceived wife of Case 3, truly believes that her husband is having an affair with Rachel. Although Nicole believes that she believes that her husband is not having such an affair, this higher-order belief is false. Since she does not believe that her husband is not having an affair, Nicole cannot let the thought that he is faithful guide her actions in any robust sense. For example, she does not let that thought guide her to Rachel's house at times that she should believe (and, indeed, *does* believe) that he is there. In a weak sense the self-deceived present behavior that is indicative of believing as they desire. They report such a belief, and utterances are behavior. But they also exhibit contrary avoidance behavior. Again, the view of belief favored here is one which places greater emphasis on non-linguistic behavioral dispositions than first-person phenomenology and linguistic reports. And the behavioral dispositions of the self-deceived, especially when in situations where the costs of mistake are high, are tipped toward believing the truth. It is unlikely that Mitchell would be willing to have a camera focus on his scalp for a shampoo commercial—there the costs are simply too high.[29] If Mitchell were to consent to such a public and critical examination of his scalp, that would be strong evidence that he is self-deluded, not self-deceived.

We can see how this higher-order account of self-deception explains the features that many have noted of self-deceivers—features that Mele's theory does not explain. The main feature that is missing in Mele's theory is the tension characteristic of self-deceivers. Audi explains it as follows:

> My positive suggestion here is that what is missing (above all) [in Mele's theory] is a certain *tension* that is ordinarily represented in self-deception by an avowal of *p* (or

> tendency to avow *p*) *coexisting* with knowledge or at least true belief that not-*p*.[30]

Audi is correct that self-deceivers do tend to say one thing, while truly believing otherwise. Private and public avowals fall under the first two abilities of higher-order beliefs. So the higher-order account explains why self-deceivers avow what they do, but nevertheless are guided in their actions by a contradictory belief. And the disharmony between the first-order and second-order beliefs clearly demonstrates a tension. For self-deceivers the first-person and third-person aspects of belief are not in accord. From the inside it seems that they believe one thing, but from the outside their behavior reveals otherwise. This tension is sometimes described as a conflict between a subconscious true belief and conscious false belief.[31] The higher-order theory also confirms this claim to the extent that higher-order beliefs correlate with consciousness and mere first-order beliefs (i.e., first-order beliefs that are not embedded as the content of a second-order belief) do not. We should add that not only must self-deceivers have a false second-order belief, they must also *lack* the true second-order belief that not-*p*. This would explain why the false claim is before consciousness, but the true claim is not. Because the self-deceived do not have contradictory beliefs at the same level, this is not a dual-belief account.

A final advantage of this second-order account is that it explains why self-deceivers cannot know of their self-deceived status.[32] Call this the *opacity of self-deception*. L. Jonathan Cohen remarks on this apparent feature of self-deception when he writes:

> Spotting self-deceit in yourself is a lot more difficult than spotting it in others, but your own self-deceit is intrinsically easier to eliminate once you have spotted it. For, once you accept that you have spotted self-deceit in yourself on some issue, it has presumably thereby ceased to exist in you on that issue.[33]

If this is not convincing, try to imagine someone remaining in a state of self-deception while knowing that they are self-deceived.

Let's return to the example of the self-deceived wife to illustrate how our second-order theory explains the opacity of self-deception. If Nicole truly believes that she is self-

deceived about whether her husband is cheating on her, then she believes that she has hidden the true belief (that he is cheating on her) from herself. So, she would believe that she believes that he is cheating on her. But on our account we have stated that Nicole lacks this second-order belief. Instead, she believes that she believes that he is *not* cheating on her. The relevant belief-states are not transparent, and this lack of self-knowledge is essential to self-deception.

The discussion in sections 1-3 suggests the following sketch of an analysis of self-deception.

> An agent is self-deceived at time *t* if and only if:
>
> 1. The agent at *t* possesses sufficient evidence to warrant a belief that not-*p*.
> 2. The agent at *t* believes that not-*p*.
> 3. However, the agent at (and since sometime before) *t* desires to believe that *p*.[34]
> 4. This desire, by prompting characteristic deceptive strategies, causes the agent to believe, at *t*, that she believes that *p*.
> 5. The agent at *t* does not believe that she believes that not-*p*.[35]

This is only a *sketch* of an analysis, because certain key concepts have been left unexplained—e.g., 'possessing sufficient evidence' and 'characteristic deceptive strategies.' The self-deceived employ the same deceptive strategies as the self-deluded, but with less success (in that they do not attain the desired belief). So, in clarifying line 4, we can borrow heavily from Mele's excellent work in identifying the deceptive tactics of the self-deluded—such as selective attention to evidence, selective acquisition of evidence, confirmation bias, etc.

While this is only a sketch of an analysis, our discussion has yielded considerable new insights into the problem of self-deception. Our section 1 discussion of the motive given in line 3 revealed that there is self-deception that is neither straight nor twisted, but indifferent. And this sketch, unlike the proposals in Mele (1997, 2001), offers a plausible explanation of cases of self-deception in our more restricted sense, as opposed to the less puzzling cases of

self-delusion. Future discussions of motivated irrationality would benefit if the deception/delusion distinction were kept in mind. The self-deceived do *not* believe as they want to believe. Instead, they possess a false belief about what they believe. This higher-order belief theory explains the tension that separates self-deception from self-delusion.[36]

Endnotes

[1] Mele (1997, 2001). The answers I defend are most in line with the view of self-deception advanced by Robert Audi (1985, 1988).

[2] Donald Davidson (1982, 1986) offers the most prominent and explicit philosophical version of the intentionalism. Quattrone and Tversky (1984) support this approach through an interpretation of their psychological experiments. Intentionalism has been recently criticized in the philosophical literature by Mele (2001) and Lazar (1999).

[3] The world-focused version has been supported by Bach (1981) and Cohen (1992). Pears (1984) and Nelkin (2002) endorse the self-focused version. Davidson (1986) and Audi (1985) are explicitly agnostic. Mele (2001) also leaves the motivational content open for cases of "garden-variety" self-deception. There are also non-desire motivational accounts that I am presently ignoring. For example, Dalgleish (1997) and Lazar (1999) consider emotion-driven motivational accounts.

[4] Mele (2001), in contrast, offers distinct accounts for straight and twisted self-deception.

[5] This notion of avoidance behavior will be prominent in our discussion. By 'avoidance behavior' I mean the sophisticated behavior of avoiding evidence that not-*p* in a way that shows the agent already possesses sufficient information that not-*p*.

[6] In this regard, the present view is compatible with Mele's psychologically well-informed account of self-deceivers as attempting to minimize costly errors (e.g., Joey falsely believing that his wife is faithful) rather than being primarily concerned with true belief. See Mele (2001), Chapter 2.

[7] See Mele (2001), Chapter 1 for a good introduction to these puzzles.

[8] Nelkin (2002), p. 396.

[9] Mele seems to simply assume that the self-deceived are successful. He does not include acquiring the belief that *p* as one of the sufficient conditions, but as a presupposition of an analysis! Before listing his sufficient conditions he states: "I suggest that the following conditions are sufficient for *entering self-deception in acquiring a belief that p*." Mele (2001), p. 50 (italics in original). Also see Nelkin (2002), p. 394.

[10] Mele (2001), Chapter 4.

[11] The analysis of self-deception in Mele (2001) is as follows: "I suggest that the following conditions are jointly sufficient for *entering self-deception in acquiring a 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 of *p* in a motivationally biased way.
3. This biased treatment is a nondeviant cause of *S*'s acquiring the belief that *p*.
4. The body of data possessed by *S* at the time provides greater warrant for ~*p* than for *p*." (pp. 50-51)

[12] Audi (1985), pp. 173-177; Audi (1988), p. 94; Audi (1997), p. 104.

[13] Bach (1997), p. 105.

[14] Martin (1997) and Perring (1997).

[15] It might be challenged that some other intentional state, besides belief, could account for Mitchell's avoidance behavior. For example, perhaps Mitchell is motivated by a fear of going bald, or by a mere suspicion that he is bald. Note, however, that a purely emotive state, such as a fear or worry of going bald, is not sufficient to account for such avoidance behavior. There must also be some belief-like state coupling with this fear to guide Mitchell in his specific behaviors. For example, Mitchell could easily have feared going bald while he

still had a full head of hair. At that time, he did not (let us plausibly assume) resort to a combover technique. This is because Mitchell lacked a belief that this fear was actualized. At this juncture, one might take a different tack and suggest that a mere suspicion of baldness, rather than full-fledged belief, would be enough to explain avoidance behavior. This possibility is not as damaging to the present view, since suspecting and believing are in the same family of intentional states. Indeed, doubting, suspecting and believing can be seen as three parts of one spectrum. Still, the sophistication and apparent well-informedness of the avoidance behavior of self-deceivers point toward attributing full-fledged true beliefs. More importantly, when the stakes become too high for self-deceivers and it becomes important that they act on their sincere beliefs, we see that self-deceivers do act in a way that shows they believe the truth. An example of this is provided in Section 3. I thank an anonymous referee for bringing these concerns to my attention.

[16] I thank an anonymous referee for bringing this complication to my attention.

[17] The deception/delusion distinction is also made in Audi (1988), p. 109: "Take a case of self-deception in avowing that one is courageous (*p*). There is a certain tension here which is characteristic of self-deception and partly explains its typical instability. The sense of evidence against *p* pulls one away from the deception and threatens to lift the veil concealing from consciousness one's knowledge that one is not courageous, but the desires or needs in which the self-deception is grounded pull against one's grasp of the evidence and threaten to block one's perception of the truth. If the first force prevails, one sees the truth plainly and is no longer deceived; if the second force prevails, one passes from self-deception to single-minded delusion and does not see the truth at all. Self-deception exists, I think, only where there is a balance between these two forces." See also Audi (1985), p. 171.

[18] In contrast, Audi (1988) presents examples of self-deception in our limited sense. (See, in particular, pp. 94-97.)

[19] Mele (2001), p. 3.

[20] Mele (2001), pp. 36-37, 32, 26.

[21] Mele (2001), pp. 79-80.

[22] The comments in Audi (1997) and Bach (1997) regarding the tension characteristic of self-deception also describe self-deception as a *process*.

[23] Audi (1985), pp. 177 and 182 suggests, but does not develop, this possibility.

[24] This requirement for the capacity of reflection explains why animals and young children cannot self-deceive. Compare with Nelkin (2002), p. 398.

[25] Rosenthal (1986, 1997) offer some of the more prominent articulations of this view.

[26] Lycan (2001) and Manson (2001), for example, see higher-order (representational) theories in this way. However, they are not discussing higher-order *thought* theories in particular.

[27] Rosenthal mentions subjects with higher-order thoughts of non-existent mental states. It should be pointed out that his higher-order thoughts are not beliefs, but his observation regarding such higher-order thoughts is still relevant. He writes:

"On the present account, conscious mental states are mental states that cause the occurrence of higher-order thoughts that one is in those mental states. And, since those higher-order thoughts are distinct from the mental states that are conscious, those thoughts can presumably occur even when the mental states that the higher-order thoughts purport to be about do not exist. But such occurrences would not constitute an objection to this account. It is reasonable to suppose that such false higher-order thoughts would be both rare and pathological." (1986), p. 338. Rosenthal holds that the higher-order thought makes the *lower order* mental state conscious. I think that the higher-order belief makes it more likely that the

lower-order belief is conscious, and when the lower-order belief is non-existent a "false consciousness" still likely occurs.

[28] Although, because even the self-deceived generally respect the weight of evidence, these first-person qualities will likely not be as constant as those of genuine believers (e.g., the self-deluded). It might take some effort for self-deceivers to maintain these first-person qualities. So, in this regard, we can concede somewhat to Bach when he writes: "The self-deceiver is disposed to think the very thing he is motivated to avoid thinking, and this is the disposition he resists." (1997), p. 105. We can say the following: In virtue of believing the truth, the self-deceived are disposed to think the truth. But in virtue of having a false higher-order belief, this disposition is overridden and the self-deceived think what they do not believe.

[29] Compare with Audi (1985), pp. 188-189. This is the example promised in footnote 15.

[30] Audi (1997) p. 104.

[31] Audi (1985), p. 173; Audi (1988), p. 94.

[32] Knowledge of deception, as I am using the terms, does not require knowing what is really true and false of the world. Rather, it requires knowing your evidence, motives, deceptive tactics, and belief states.

[33] Cohen (1992), p. 147.

[34] This motive might be narrowed to a desire merely for the first-person aspects associated with believing that *p*, as discussed earlier.

[35] Many would want to include a sixth clause stating that not-*p* is true. I do not think this is necessary, however. The tension that is symptomatic of self-deception can certainly arise without this requirement. But, if one thinks that as a matter of proper usage deception requires a motivation to believe what is false, they can simply add such a sixth clause.

[36] Thanks to Steffen Borge, Tamar Gendler, John Hawthorne, Karson Kovakovich, and Ted Sider for helpful comments on an earlier version of this paper.

REFERENCES

Audi, Robert. (1985). "Self-Deception and Rationality," in *Self-Deception and Self-Understanding*, ed. Mike W. Martin (Lawrence: University of Kansas), pp. 169-194.

Audi, Robert. (1988). "Self-Deception, Rationalization, and Reasons for Acting," in *Perspectives on Self-Deception*, eds. Brian McLaughlin and Amelie Rorty (Berkeley: University of California Press), pp. 92-120.

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

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

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

Cohen, L. Jonathan. (1992). *An Essay on Belief and Acceptance*, (Oxford: Clarendon Press).

Dalgleish, Tim. (1997). "Once More with Feeling: The Role of Emotion in Self-Deception," *Behavioral and Brain Sciences* 20, pp. 110-111.

Davidson, Donald. (1982). "Paradoxes of Irrationality," from *Philosophical Essays on Freud*, Richard Wollheim and James Hopkins, eds., (New York: Cambridge University Press), pp. 289-305.

Davidson, Donald. (1986). "Deception and Division," from *The Multiple Self*, Jon Elster, ed., New York: Cambridge University Press.

Lazar, Ariela. (1999). "Deceiving Oneself or Self-Deceived? On the Formation of Beliefs 'Under the Influence'," *Mind* 108, pp. 265-290.

Lycan, William. (2001). "A Simple Argument for a Higher-order Representation Theory of Consciousness," *Analysis* 61, pp. 3-4.

Manson, Neil C. (2001). "The Limitations and Costs of Lycan's 'Simple' Argument," *Analysis* 61.4, pp. 319-323.

Martin, Mike W. (1997). "Self-Deceiving Intentions," *Behavioral and Brain Sciences* 20, pp. 122-123.

Mele, Alfred. (1997). "Real Self-Deception," *Behavioral and Brain Sciences* 20, pp. 91-102.

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

Nelkin, Dana. (2002). "Self-Deception, Motivation, and the Desire to Believe," *Pacific Philosophical Quarterly* 83.4, pp. 384-406.

Pears, David. (1984). *Motivated Irrationality,* (Oxford: Clarendon Press).

Perring, Christian. (1997). "Direct, Fully Intentional Self-Deception is also Real," *Behavioral and Brain Sciences* 20, pp. 123-124.

Quattrone, G., and A. Tversky. (1984). "Causal versus Diagnostic Contingencies: On Self-Deception and on the Voter's Illusion," *Journal of Personality and Social Psychology* 46, pp. 237-248.

Rosenthal, David. (1986). "Two Concepts of Consciousness," *Philosophical Studies* 49, pp. 329-359.

Rosenthal, David. (1997). "A Theory of Consciousness," reprinted in *The Nature of Consciousness*, eds. N. Block, O. Flanagan, and G. Guzeldere, (Cambridge, MA: The MIT Press).