Responsibility and Self-Deception: A Framework

Dana Kay Nelkin *
dnelkin@ucsd.edu

ABSTRACT

This paper focuses on the question of whether and, if so, when people can be responsible for their self-deception and its consequences. On Intentionalist accounts, self-deceivers intentionally deceive themselves, and it is easy to see how they can be responsible. On Motivationist accounts, in contrast, self-deception is a motivated, but not intentional, and possibly unconscious process, making it more difficult to see how self-deceivers could be responsible. I argue that a particular Motivationist account, the Desire to Believe account, together with other resources, best explains how there can be culpable self-deception. In the process, I also show how self-deception is a good test case for deciding important questions about the nature of moral responsibility.

Introduction

Self-deception is a phenomenon that manages to strike us as both very common and yet not easy to characterize. It is easy to see how one person can deceive another; after all, one person knows the truth and, using any of a variety of techniques, can influence another to believe a falsehood. But how can one person deceive herself?

Some have decided that self-deception is impossible, but most theorists writing on the subject have continued to assume that it is a real phenomenon, while acknowledging that the correct model of self-deception must diverge in at least some ways from that of interpersonal deception. For the last several decades, those working in the area have tended to occupy one of two main positions on the question of what self-deception is: Intentionalism and Motivationism. Intentionalists preserve at least two key components from the

* University of California, San Diego, USA.

model of interpersonal deception, arguing that self-deception is intentional and that the self-deceived person holds a true belief while at the same time wrongly believes the contrary. In order to address the question of how one person who knows the truth could possibly convince herself of its contrary, the Intentionalist picture is most often combined with the view that a single person can be divided, or partitioned in some way, so that one part of her believes the truth, and brings about the contrary belief in the other part of her.[1] Such a picture allows us to think of self-deception as very like interpersonal deception while also maintaining a great deal of explanatory power. It helps explain why it is that self-deception is a sophisticated cognitive activity, not possessed by young children, for example. It also helps explain why we often hold self-deceivers responsible for their deception and for the consequences that follow from it. It can account for a wide variety of cases of self-deception, including cases in which self-deceivers believe things about the world that they want to be the case, as well as cases of so-called "twisted" self-deception, in which they believe things about the world that they would rather not be true. (For example, consider the case of a husband who desperately wants his wife not to be having an affair, but, worried about getting caught off guard, convinces himself that she is.) All that is required to account for both cases is that self-deceivers intentionally engage in the formation of a belief they know to be false.

Despite these theoretical virtues, and the fact that Intentionalism may once have been the dominant picture, it has lost ground in recent years to Motivationists. They reject Intentionalism on the grounds that it either leads to paradox or, at a minimum, to the unnecessary and unsupported postulation of strongly autonomous parts of the mind.[2] Opponents of Intentionalism (myself included) claim to be able to capture (most of) its theoretical advantages without the metaphysical and psychological complexity of partitioning.

Motivationists have in common the commitment to the idea that self-deception involves a kind of motivated state, while rejecting the commitment to the deception being an intentional action. Beyond this common commitment, motivationists divide in their answers to several further questions about the nature of self-deception, including these:

[1] See Pears 1984, for example.
[2] See Mele 1987 and Johnston 1988 for examples.

1) What is the guiding motivation? Assuming it is a desire, a desire for what? (the content question)
2) What is the product of self-deception about p? A belief that p? A sincere avowal that p? A pretence that p? A belief that one believes p? (the product question)
3) How does the motivation generate the product of self-deception? (the process question)
4) What accounts for the irrationality in self-deception? (the irrationality question)
5) Is there a belief that not-p? (the contrary proposition question)
6) If the product of self-deception is a belief that p, must that belief be false? (the truth value question)

To see the disagreements starkly, it helps to consider some examples. Some argue that the product of self-deception is a false belief that *p* and that there is no contradictory true belief, while others argue that self-deception requires a true belief that *not-p*, and no contradictory false belief that *p*. It is puzzling how a self-deceiver could have no false belief about which she is self-deceived, but various alternatives are suggested in its place in answer to the product question. For example, there are those who argue that rather than having a false belief, the self-deceiver sincerely avows a false claim (the avowal view)[3]; those who argue that the self-deceiver pretends that the false belief is true (the pretence view)[4]; those who argue that the believer has a false belief about her own states of mind, rather than about the object of self-deception (the failure of self-knowledge view)[5]; those who argue that the believer acts in some ways as if she believes the false belief (the behavior view)[6] and combinations thereof.

Notably, motivationists also divide on the question of the nature of the motivation in question. Must it be a desire or could other emotions, such as anxiety suffice?[7] If self-deception must be driven by a desire, what is the content of that desire?[8]

[3] See, for example, Audi 1997 and Funkhouser 2005.
[4] For example, Gendler 2007.
[5] See, for example, Scott-Kakures 1996, Fernandez 2011, and Funkhouser 2005.
[6] See, for example, Audi 1997 and Funkhouser 2005.
[7] See, for example, Barnes 1997.
[8] See, for example, Mele (1997, 2000, 2001) for an account in which the desire need have no particular content, and Mele's earlier (1987) for an account in which the desire must be the desire that p be true, where the product of self-deception is the belief that p. See Nelkin 2002 for a different account to be explained shortly.

The best account of self-deception will answer all of these questions in a plausible and coherent way, as well as yield explanations of some important and well-recognized features of self-deception. Ideally, the account would explain why self-deception takes considerable cognitive sophistication (and thus, why young children do not seem to be capable of it), why we often attribute responsibility on the part of self-deceivers for their self-deception and its consequences, and why we have the thought that self-deception shares similarities with other-deception.

In this paper, I will focus centrally on one challenge that faces all motivationists, and in doing so bring out the virtues of one particular kind of motivationist account. One of the claimed virtues of intentionalism is that it can explain why we often hold self-deceivers responsible and, indeed, often blame them for their deception and for the consequences that follow. If self-deception is an intentional act, then it is chosen by the agent herself, in full knowledge of what she is doing, and is a paradigm object of blame. But if we leave intentions out of the picture, and we think that self-deception is paradigmatically an *un*intentional (and often unconscious) process, then it becomes less obvious that the self-deceived are responsible for their states and for their consequences.

In an earlier paper, I argued for a particular motivationist account, the Desire to Believe account, that seemed most naturally to explain the fact that we often hold people responsible for their self-deception.[9] I believe that the account is especially well-suited for this task, but I also believe that there are more questions to be raised for motivationist accounts, including the one I defend. And in this paper, I articulate these questions, and develop answers based on the Desire to Believe Account. It is worth noting that some (but not all) of what I say could also be adopted by other motivationist accounts. I begin in section 2 by setting out the motivationist account I favor. In sections 3 and 4, I then elaborate some challenges to accounting for responsibility, homing in on where the important issues lie. While I will not here defend a comprehensive theory of responsibility, I will make a start in identifying the issues that must be resolved in attributing responsibility for self-deception, as well as locating self-deception relative to other sorts of objects of responsibility in which we have confidence in our attributions of responsibility (or non-

[9] See Nelkin 2002.

responsibility). In the process, I also hope to show how self-deception can provide an illuminating test case for certain theories of responsibility.

1. The Desire to Believe Account

It will be helpful to lay out some preliminary methodological assumptions. First, I begin with the assumption that the problems with the intentionalist picture are difficult ones, and that if a motivationist picture can succeed in accounting for cases of self-deception and can explain the phenomena we think need to be explained, then we should adopt it.

Second, an account that offers a set of necessary and sufficient conditions would be useful and neat, but it may be that there are blurry boundaries, and that there isn't a perfectly neat set of necessary and sufficient conditions. In that case, it would be better to adopt an account of conditions that are sufficient for self-deception, and that also characterize all the central cases of self-deception, distinguishing it clearly from phenomena that we believe are distinct (such as certain "cold" or unmotivated kinds of belief formation like simple cognitive error, as well as other sorts of "hot" or motivated belief formation such as wishful thinking.)

Endorsing this approach, I aim to combine insights of both intentionalist and motivationist models in order to arrive at a model for being self-deceived about a proposition, say, *p*. The model offers a sufficient condition for being self-deceived, and, I believe, necessary conditions for all the central cases of self-deception that distinguish it from other well-recognized phenomena. The key and distinctive aspect of the account is its answer to the content question: the guiding motivation in self-deception about *p* is a desire to believe that *p*. Accounts that leave open the content of the desire in question, such as Mele's (1997), are appealing in their flexibility, but they also suffer from failing to capture what distinguishes self-deception from other sorts of "hot" belief formation. For example, consider the case of Otis who is motivated to have an answer to every question. When asked whether the 1991 Braves would have prevailed over the 1999 Yankees, his desire to have an opinion motivates him to focus on a particular set of statistics (while ignoring others that might induce doubts) allowing him to form the view that the Braves would have won. Intuitively, this is not a case of self-deception. Or consider the case of Ben, a small child who comes to believe, against his evidence, that his babysitter is not a nice person. He desires his parents' return, and through a complex defense

mechanism forms this belief as a result.[10] These are cases of "hot" or motivated biased belief formation, but do not seem to be ones of self-deception. The category of motivated and biased belief seems, intuitively, larger than that of self-deception.

Traditionally, accounts of self-deception that are more restrictive in the content of the desire in question have tended to identify the relevant desire as the desire that *p*.[11] But, as Mele (1997) points out, this excludes clear cases of self-deception, namely, those of the "twisted" variety. We can see, however, that the desire to *believe* that *p* is plausibly attributed in both paradigmatic straight and twisted cases, and nicely excludes cases like that of Otis and Ben and the Babysitter and other cases of motivated belief that are not self-deception. This answer to the content question also shows why the intentionalist view is appealing, even though false: intentionalist views take it that the intention to deceive either arises from, or is partly constituted by, a desire to generate a belief that *p*. This aspect of the intentionalist view is thereby preserved.

How does the view answer the other questions? As for the product of self-deception, I believe that the most natural answer is that it is a belief. If it is, then once again self-deception will retain a key feature of deception in general. Further, taking it to be a belief explains the (several) kinds of behaviors that the self-deceived person then engages in.[12] To take an example, the mother who is self-deceived in believing her son will return keeps his bedroom undisturbed, sincerely swears that he will return, and makes sure that the house numbers are always lighted. Her behavior is well explained by her believing that he will return.

Finally, in conjunction with identifying the desire to believe as the content of the guiding motivation, understanding the product of self-deception as a belief gives self-deception a kind of intelligibility that is otherwise lacking. *Were* the agent to become aware of her desire to believe, she would be able to see immediately that the product satisfied her desire. No special knowledge of

[10] I discuss these cases in more detail in Nelkin 2002.

[11] See, for example, Mele 1987.

[12] As mentioned, there have recently been a number of alternative suggestions made as to what the product of self-deception is. See notes 3-6. Some also argue against the claim that belief is the product on the grounds that certain behaviors of the self-deceiver, such as the avoidance of evidence, does not fit well with it. But I believe that this behavior can be well accommodated by understanding it as motivated treatment of the evidence.

defense mechanisms and their function would be required to see that her desire had been satisfied.[13]

As for the process, I believe that in its details this is an empirical question, and there may be a great variety of ways that the motivating desire operates to result in the belief that *p*. But we can say some things very generally about it. It seems that the desire has an influence on the agent's treatment of the evidence available to her – either in the selection of data she focuses on, or in the inferences she draws from it – so that she sees the evidence as supporting the belief that *p*, even though it in fact provides greater support for *not-p*.[14]

This answer to the process question also leads naturally to the question of where the irrationality is to be found. It need not be in starkly contradictory

[13] Mele (2009) has argued recently that the key aspect of my account that underlies this intelligibility claim is incorrect. In particular, he offers a counterexample to the claim that a desire to believe is necessary for self-deception. He asks us to «imagine two jealous husbands with very similar evidence in very similar circumstances. Each acquires the false, unwarranted belief that his wife is having an affair--the belief that a, for short» (2009, p. 268). Both treat the evidence they have in similar biased ways. One husband, Jack, is motivated by the desire to believe that a. The other, John, lacks that particular desire, but «does have desires that contribute to his having acceptance and rejection thresholds for a that are just like Jack's. Suppose for good measure, that John has a desire not to acquire a false belief that his wife is innocent of infidelity» and that this desire is what motivates his biased treatment of the evidence and acquisition of the belief that a (p. 268). Because the two husbands are so similar, «it is very plausible that if Jack is self-deceived, so is John» (p. 269). I do not believe that these cases give us good reason to reject the Desire to Believe account. First, as mentioned, the account is consistent with there being blurry boundaries between self-deception and other sorts of irrational motivated belief formation. But there are boundaries, nonetheless, and it is crucial to evaluating the example to distinguish between John's "having desires that contribute" to the biasing as Jack does and his having the very particular desire that Mele offers us "for good measure." Cases like Otis and Ben and the Babysitter show us that restriction on the content of desires is essential to distinguish self-deception from other sorts of cases, and to my knowledge, Mele does not respond to this concern. Equally importantly, the case of John is one in which the content of the motivating desire is actually very similar to that of Jack's. It is a desire not to have a false belief that p, rather than to acquire a belief that not-p. Thus, I believe that Jack's case falls at best in the blurry boundary area of self-deception. This is precisely because the content is so similar as to retain something approaching the intelligibility provided by the Desire To Believe account. (Both Jack's and John's desires have as their objects beliefs with 'p' embedded in its content as the relevant object of desire.) And yet the content is not so similar as to be as easy to see that a desire is fulfilled by successful self-deception; and the content is different enough that it fails to capture the key similarity to other-deception that the account provides. All other things equal, the closer the content of the desire gets to the belief that not-p, the more clearly it is a case of self-deception, on the Desire to Believe account, and, conversely, the farther the content gets from the belief that not-p, the less clearly it is a case of self-deception. If, as I have argued, we confine ourselves to the most specific version Mele offers of John's situation, these cases do not appear to give us reason to doubt this.

[14] See Kunda 1990 for a classic statement of a theory of motivated belief.

beliefs, as it is on the intentionalist picture, but on this motivationist picture, we find it in the logical tension between the self-deceptive belief and the agent's evidence. This in turn can help explain the seemingly odd behaviors that are found in many paradigmatic cases of self-deception. On the one hand, the product of self-deception is the belief that *p*, but there is also avoidance of evidence that better supports *not-p*, for example. Both of these behaviors can be explained by appeal to this sort of mechanism of motivated selective evidence gathering.

On the Desire to Believe account, it is not necessary to have a belief that *not-p*. It is consistent with the account that some cases of self-deception include such a belief, and therefore that some self-deceivers do have contradictory beliefs, but it is not required, as long as the other conditions are met.[15]

How does the account answer the truth value question? I now believe that the account can be open on this question without loss of explanatory power. Typically, we assume that if someone is deceived, she believes something that is false. But by coincidence, it could turn out that though her evidence fully supported *not-p*, and her desire to believe *p* led her, via a biased treatment of evidence, to believe that *p*, she got lucky (so to speak) and *p* is true. Since the psychological process is the same regardless of the truth value of *p*, I think it is reasonable to treat even such a "lucky" case as one of self-deception. But if one prefers instead to treat such a case as very like self-deception, but not *strictly* self-deception, I see no objection to doing so.

Putting the pieces together then, we have a view according to which a person is self-deceived with respect to *p* when she believes that *p* as a result of the biased treatment of evidence which is in turn motivated by the desire to believe that *p*, and when the evidence available to the person better supports *not-p*. In addition to the advantages already described, the view succeeds in explaining why a certain degree of cognitive sophistication is required for self-deception, since it requires having a desire to believe – a second order desire

[15] Some have argued that "at some level" one must believe the truth, e.g., Funkhauser (2005) and Audi (1997). If this turned out to be true, it could be added as a condition to the account. And the kind of partitioning it would require would still be substantially less robust than that of the intentionalist picture in which the mind is divided into true sub-agents. I think it is psychologically plausible that we do have such partitioning of beliefs in some instances. But I do not think it is necessary to account for all of the features of self-deception. The behavior that it is thought to be essential to explain seems to me to be explainable simply by the strong desire to believe that p, operating through a mechanism of selective evidence gathering.

about one's own mental states. This is a significant advantage of the view over accounts that leave the content of the guiding desire unrestricted. And as we have seen, it preserves several aspects of the intentionalist picture, showing why, even if that picture is incorrect, it has been regarded as attractive.

2. Motivationism, Self-Deception, and Responsibility: First Pass

With this particular motivationist account in view, we can turn to the question of whether, and, if so, when self-deceivers are responsible for their self-deception and its consequences.

One answer to the question of when one might be responsible for self-deception is that no one ever is, on the grounds that no one is responsible for anything. This position on responsible action has been forcefully defended and has a number of adherents.[16] I will here set aside this challenge, however, and concentrate on reasons for thinking that the move to motivationism causes a special reason for skepticism.

I will here adopt a very general approach to moral responsibility that understands responsibility to depend on control.[17] In particular, for an agent to be responsible for her actions, she must be responsive to reasons. There are a number of ways that this idea has been developed, and we will see that the details may matter when it comes to judging the responsibility for self-deception. For example, some have defended mechanism-based approaches to reasons-responsiveness, arguing that the responsible agent must act on a mechanism that is reasons-responsive, while others have defended agent-based approaches, arguing that it is the responsible agent herself that must be responsive to reasons.[18] For now, let us begin with the intuitive idea that to be

[16] See, for example, Pereboom 2001.

[17] I here sidestep the important issue of whether moral responsibility is consistent with determinism, for the reason that I want to concentrate on the particular question of whether motivationism has special difficulties accounting for it that intentionalism does not. It is worth noting, however, that incompatibilists–those who deny that responsibility is consistent with determinism–often accept conditions like those described here as necessary, even if not sufficient. (For example, see O'Connor 2001.)

[18] The most notable defenders of a mechanism-based approach are Fischer and Ravizza, who argue that to be responsible one must act on a moderately reasons-responsive mechanism for which one has taken responsibility (2000). Levy (2004) adopts this account in discussing self-deception. Defending an agent-based approach, Wolf (1991) argues that the responsible agent must be able to act on the reasons there are. Some, like Wallace (1994), also defend an agent-based approach, and combine this with the claim that a responsible agent must have general rational capacities. In contrast, others,

responsible, or *a fortiori*, blameworthy for one's actions, one must be able to respond to the reasons that there are.

Questions about the relationship between specific models of self-deception and responsibility have been raised continuously throughout the contemporary discussion of the phenomenon. For example, in a classic article advocating an intentionalist picture of self-deception, Demos writes: «A man who lies to himself is blameworthy because he acts with knowledge of the facts and thus may be held responsible for his erroneous belief» (1960, p. 589). Writing a few years later, Fingarette seems to draw a quite different conclusion:

> There is thus in self-deception a genuine subversion of personal agency and, for this reason in turn, a subversion of moral capacity. The sensitive and thoughtful observer, when viewing the matter this way, is inclined not to hold the self-deceiver responsible but to view him as a 'victim'. (Fingarette, 1969, p. 140.)

And writing recently on the topic, Neil Levy suggests that the motivationist conception of self-deception in particular «has neither need nor place for attributions of moral responsibility to the self-deceived in paradigmatic cases» (2004, p. 294).[19] Is there reason to think that motivationists, in particular, have difficulty accounting for moral responsibility in paradigmatic cases?

The argument that most cleanly distinguishes cases satisfying intentionalist conditions from those satisfying motivationist ones is based on the claim that we can only be responsible where there is intentional action. Choice, it has been argued, is the locus of responsibility, and it does not make sense to hold people responsible for things that they did not choose.[20]

But this view is undermined by a number of apparent counterexamples, and we do not have to look to the hard cases of self-deception to find them. Cases of recklessness seem to qualify as responsible actions. For example, it has happened that though people do not intend to create a risk to their neighbors,

including Fischer and Ravizza and Wolf, argue that the relevant abilities must in some sense be exercisable in the particular situations in which they are responsible. I develop and defend a view that is agent-based and that requires abilities to exercise rational powers in particular situations in Nelkin 2011.

[19] This strong claim appears to be tempered by other claims in Levy's paper that make his view somewhat difficult to pin down. In the conclusion of his paper, for example, he acknowledges that self-deceivers may "often" be responsible (2004, p. 310). But setting this aside this uncertainty, his paper makes explicit the important question of just how much and when self-deceivers are responsible. See also the interesting treatment of Linehane (1982) that also focuses attention centrally on this issue.

[20] See Alexander & Ferzan (2009) for an articulation of this theory.

they consciously disregard knowledge of the risk in setting off firecrackers in the street, for example.[21] Despite not harming or even creating risks intentionally, they are blameworthy for acting as they do. This can be explained by their having had the ability to respond to reasons – an ability that they failed to exercise.

Still, even if we do not draw a line around intentional action, there have been tempting reasons for drawing it in another place that also excludes culpable self-deception on the motivationist picture, namely, around awareness. That is, it has been argued that people cannot be responsible for acting in harmful ways (or unreasonable-risk-enhancing ways) if they are unaware of the risks in question.[22] For example, if I am completely unaware that my light switch has been hooked up to a stick of dynamite currently located in my neighbor's kitchen and I flip the switch, thereby causing the dynamite to explode, I am not blameworthy for anything I did. Having no access to the relevant reasons, I couldn't have responded to them. Here, too, intentionalism and motivationism seem to fare differently. If the self-deceiver intentionally deceives, then it simply follows that she knows of the risk in question. (In fact, she is trying to maximize it.) But if self-deception is an unintentional, and likely unconscious process, then she need not be aware at all of a risk of generating a self-deceptive belief.

But this requirement on responsible action also appears to be too strong. There are cases in which one is unaware of the risk one creates (or fails to stop), but it remains the case that one *ought* to have been aware of it, and so is blameworthy for failing to acquire awareness and act accordingly. Again, we need not look to self-deception itself to see that this is the case. A person who tells an offensive joke may not be aware of the risk that his audience will take offense, but might very well be blameworthy for having done it and responsible for the effects.[23] Thus, if the criterion for responsibility is not awareness, but rather that one "should have known", *and* if self-deception at least sometimes satisfies this criterion, then it is possible to see self-deception as a case in which people are sometimes responsible. Exactly how often and when will then depend at least in part on whether the self-deceiver satisfies the "should have

[21] See the Model Penal Code on recklessness and Alexander and Ferzan (2009, p. 25).
[22] This is what Sher (2009) calls the "Searchlight View".
[23] For this and related cases, see Sher (2009, p. 28).

known" condition, and, of course, the conditions under which one "should have known."[24]

How we are to understand these conditions is itself a matter of fairly intense controversy. But before entering into that debate, I think it is already possible to see how the Desire to Believe Account is in one way better suited than other motivationist accounts to show that such conditions are at least sometimes satisfied.

On an *unrestricted* desire account, it is admittedly not obvious how a self-deceiver could be responsible. For example, if the product of self-deception is the belief that *p*, and the guiding desire is a desire for *q*, then the operative mechanism might be one that is quite complex and that we could not expect the self-deceiver to be aware of. For example, citing research on jealousy, Mele (2001) suggests that a case of self-deception might have stemmed from a desire to have closer relationships, and through a protective mechanism lead to a belief that one's spouse is having an affair. If the content of the desire is so far removed from the content of the product of self-deception, it seems unreasonable to expect the self-deceiver to have even been on guard against such a process.[25] But cases that satisfy the conditions of the Desire to Believe Account have a kind of immediate intelligibility that these cases do not. Were the agent to be aware of her desire to believe that *p*, she could immediately see that her belief that *p* satisfies her desire. This is not yet to say that she is responsible for her deception; but it does make clear how it could be comparatively easier for her to be on guard not to form this kind of motivated belief against the evidence.[26]

Of course, it is open to the advocate of an unrestricted account to allow that some cases satisfy the narrower conditions of the Desire to Believe account, and given that the account only aims to give sufficient conditions, can allow even that some cases are intentional (though that would admittedly require allowing for the strong partitioning that there is reason to be skeptical of). Still,

[24] It is also possible that in some cases of self-deception, people are aware of the risk in a general way, even if not of the specific process.

[25] For further discussion of this case, see Nelkin 2002.

[26] In this way, the Desire to Believe account already anticipates one line of criticism Levy (2004) levels against motivationist accounts that try to preserve the claim that (some) paradigmatic instances of self-deception are one for which people are responsible. He there criticizes other motivationist views precisely on the basis of the differential content between desire and belief (2004, p. 308), asking how one could possibly see the relationship between them. But this criticism does not apply to the Desire to Believe account.

thinking of cases with the desire to believe as *central* allows for comparatively more attributions of responsibility.

At the same time, it is important to note that the Desire to Believe account does not entail that all cases are ones for which one is responsible either. This flexibility is welcome, I believe. After all, figuring out what we are responsible for – if anything – in the way of actions, omissions, and states of all kinds has itself been the source of enormous controversy, and at least some of this controversy rests on debates about the empirical facts concerning the capacities of human beings. We should be cautious in approaching the question of how often, if at all, people are responsible for their self-deception and its consequences. So far, then, I am making a simple comparative claim: the Desire to Believe account has an advantage over other motivationist accounts in that the surface intelligibility of the relationship between guiding motivation and belief makes it easier, all other things being equal, to either be aware of the non-rational process of belief formation, or at least to be on guard against it.

3. Developing a Framework

If what I have argued so far is correct, then the Desire to Believe account has one advantage over other motivationist accounts in accounting for responsibility. But this leaves open all sorts of questions, including the conditions under which any particular instance of self-deception is something for which the self-deceiver is responsible. In this section, I will spell out some issues whose resolution is needed to make progress in answering these questions, and distinguish two sorts of strategies we can take toward making particular determinations of responsibility. I will conclude by showing how each can work.

First, it is important to get clear about exactly what the self-deceiver is supposed to be responsible for. We should distinguish between the process of self-deception, the immediate product of self-deception, and its more indirect consequences. The point made so far on behalf of the Desire to Believe account addresses the process directly, as the account shows how a person could more easily be on guard to avoid the process itself given its surface intelligibility. But even if one had no knowledge of the process, or even any reason to be on guard against it, one might still be responsible for the product. How can this be? Suppose, as is plausible, that one has a duty to form beliefs

that conform to the available evidence, particularly in cases in which much is at stake. Then even if one is not (and has no reason to be) aware of the self-deceptive process that generates a belief undermined by the evidence, it can still be that one ought to critically examine such beliefs and eliminate them. Alternatively, it can be that one ought to engage in simultaneous processes, which would compete with the self-deceptive one by including seeking out and carefully evaluating relevant evidence. These latter points are consistent with a variety of motivationist accounts of responsibility, not just the Desire to Believe account.

These points also illuminate the path to an insight about the nature of responsibility itself. They do so by supporting one general way of developing the reasons-responsiveness approach to responsibility. To see how, recall that there are different ways that the approach has been developed: we can require that the responsible agent act on a *mechanism* that is itself responsive to reasons, or we can require that the responsible agent *herself* be reasons-responsive in the relevant circumstances.[27] On the former view, if the motivated biasing mechanism of self-deception is not reasons-responsive (as seems plausible), then the self-deceiver will not be responsible. In contrast, on the latter view, it is not exonerating that a non-reasons-responsive mechanism is operating. What matters is whether *the agent* could have either prevented that mechanism from operating, or instead put another into action. Since there is good intuitive support for the idea that self-deceivers can be responsible, the agent-based approach to responsibility and the motivationist picture of self-deception provide each other with mutual support.[28]

Finally, in figuring out what agents are responsible for, we should note that it is not easy to say what the relationship is between our responsibility for our

[27] See note 18.

[28] I agree here with one part of DeWeese Boyd's (2010) discussion of Levy's application of a mechanism-based account. He writes: «However, the question isn't whether the biasing mechanism itself is reasons responsive but whether the mechanism governing its operation is, that is, whether self-deceivers typically could recognize and respond to moral and non-moral reasons to resist the influence of their desires and emotions and instead exercise special scrutiny of the belief in question» (2010, section 5.1.). Where I part company is that I reject the necessary attribution of reasons responsiveness to any mechanism--whether the biasing one or the governing one. (I am actually unsure how one would individuate the biasing mechanism and the mechanism "governing its operation".) Thus, I also reject the equivalence of reasons-responsiveness of governing mechanisms of biasing mechanisms with that of agents. But I agree with the second half of the equivalence claim, namely, that what matters for responsibility is what the agent's abilities are. Further, I claim that the case of self-deception actually helps adjudicate between mechanism-based and agent-based views.

actions and attitudes on the one hand, and their consequences on the other in general. Even where self-deception is not at issue, this is not obvious. Must the consequences be foreseeable? Must they be foreseeable to have a high probability of occurring, given one's actions or attitudes? How high must the probability be? These are difficult questions to answer. Nevertheless, it seems reasonable to conclude that where the consequences are fairly obvious results of one's actions or judgments, we have a better *prima facie* case for one's being responsible for the consequences if one is responsible for the action or attitude.

This point leads naturally into a second fundamental issue that we must address in determining responsibility. Although it might be possible to make some generalizations about sets of cases, instances of self-deception vary on a number of dimensions that can be independently relevant to responsibility (and to the degree of responsibility). For example, as we've just seen, what could reasonably be expected in terms of perceived risk plays a role in determining responsibility for consequences generally. And this would seem to apply to self-deception no less than to other cases of responsible action or omission. Suppose the self-deceived husband couldn't have possibly predicted that his wife would attempt suicide on learning of his belief that she is having an affair. Then the extent of his responsibility for such a consequence and of his blameworthiness for his judgment will depend on the risks as he could reasonably understand them at the time.[29] This case shows that there is a second factor at work in addition to degree of risk and that is severity of the harm risked, as well.

In other words, what is at stake in forming the self-deceptive belief plays a role in determining blameworthiness. A case that brings this out perhaps even more starkly is a case of a parent who is self-deceived in believing that her child is not abusing his own children. The failure to treat the evidence appropriately in this case results in her not reporting her child or protecting her grandchildren from further abuse.[30] The high stakes are not by themselves sufficient for attributing a high degree of blameworthiness, but they are one

[29] I here set aside the very large question of moral luck in consequences of one's actions. I instead assume that one is responsible for the consequences of one's actions, whatever they are, but that one's degree of blameworthiness depends not on the actual consequences, but on what it was reasonable to expect in terms of perceived risk. Thus, if the risk were high that his wife would attempt suicide and could easily be discerned, but she did not in fact attempt it, he would be equally blameworthy as in the situation in which she did.

[30] For a similar kind of case, see Barnes 1997.

factor that could potentially distinguish the case from another, like it in other ways, save for the fact that the stakes are not so high. Interestingly, Levy does not consider cases of this sort in arguing that motivationists ought to abandon the claim that self-deceivers are responsible in (some) paradigmatic cases. But it is precisely in cases like this that we can see the cost of abandoning what is an intuitively powerful thesis. Even if it were a cost ultimately to be borne, cases like this show how significant it is, and how strong the arguments would have to be for paying it.[31]

The flip side of high stakes, understood as harms to others, is the potential cost to oneself in avoiding the self-deception. This, too, is a factor that might vary from case to case. If a person could simply not go on living once having acknowledged the truth about his spouse's fidelity, for example, that seems a different sort of case than a case of someone for whom recognizing the truth would merely cause some feelings of embarrassment. A related but separable factor is the simple level of difficulty required to avoid the self-deception or to eliminate its product. There is likely a strong correlation between high stakes and difficulty, but it is possible that even where the stakes are not so high, it might, for some other reason be difficult to avoid.

Taking these two points together, we see that in making specific attributions of responsibility, we will need to take each case on its own terms and distinguish between process, product, and consequences on the one hand, and specific features of the case relating to the stakes and level of difficulty on the other. This suggests taking a fairly individualized approach to particular cases.

Yet at the same time, there may be ways of grouping certain sorts of cases together once we understand better the conditions for responsible negligence. As mentioned in the last section, this is itself a matter of great debate among both philosophers and legal theorists.

The question of how we should treat negligence – or, as Alexander and Ferzan put it, «inadvertent creation of unreasonable risks» – is not settled (2009, p. 69). And although there are relatively few skeptics about culpable

[31] Linehan (1982) and Jenni (2003) offer powerful examples of reported mental lives of Nazi doctors who collaborated–unknowingly?–in the deaths of hundreds of innocent people. While we would need to know a great deal about the cases to determine whether they were genuinely self-deceived about what they were doing, the fact that self-deception is even a relevant hypothesis in explaining their behavior provides a good example both of how much can be at stake and the intuition that self-deceivers can be responsible.

negligence, there is also much debate about *how* one can be responsible for something of which one is unaware, when faced with skeptical arguments.[32] The question is how we can be responsible for not knowing something, or not recognizing it, when we are unaware precisely of what we are supposed to know. In a recent article on the subject, Moore and Hurd (2011) reject skepticism, sticking with their «strong, bottom line intuition that blame can rightly be attached to many of the examples of negligent conduct» that appear in courts of law. But they also conclude that they are unable to come up with a «new, unifying theory of why negligence is culpable», settling for a disjunction of several sorts of conditions. (Moore & Hurd, 2011, pp. 191-192).

In light of the unsettled nature of the debate about culpable negligence in general, there are two strategies we can take to self-deception in particular. The first is to defend a general theory of negligence, and then apply it to a range of cases of self-deception. The second is to leave open what the full theory of negligence is, and instead take the more modest approach of comparing cases of self-deception to other sorts of negligence about which we have some confidence. Let us take each in turn to see how it might be developed.

One general theory of how people can be responsible for negligence and its consequences is a kind of "tracing" account.[33] According to this sort of view, one might be responsible now for one's self-deceptive belief even though one is completely unaware that it is self-deceptive, as long as one's belief is due to an earlier moment of choice during which one was aware of the risk of biased belief, could have chosen to put obstacles in the way, and did not. At the earlier time, one was reasons-responsive, and one chose badly. One's responsibility for the later deception and further consequences "traces back" to that moment. While I do not think that this covers all cases that intuitively count as culpable negligence, it might very well account for a significant number of cases of culpable self-deception.[34] It may be that there are moments in typical

[32] For two excellent recent treatments, see Moore & Hurd (2011) and Sher (2009).

[33] See Sher (2009) for a discussion of this kind of view.

[34] Levy considers this sort of account, but quickly moves on, on the grounds that it only explains why self-deceivers are "sometimes" responsible (2004, p. 304). He claims that motivationists make the "much stronger" claim that "at least typically self-deceivers are culpable" (2004, p. 304) and that they retain the "presumption" that self-deceivers are responsible. (2004, p. 310). The various claims lead to the question of what it is that motivationists have actually claimed. On my reading, at least some motivationists make nothing so strong as the claim that responsibility is a presumption in cases of self-

cases of self-deception when self-deceivers are aware of a choice to look into a piece of evidence more systematically, for example, and they choose not to, thereby allowing the process of self-deception to continue. Ultimately, it is an empirical question what sort of awareness accompanies any particular instance of self-deception. Far from being a disadvantage for the account, though, I take this flexibility to be an advantage for the reasons spelled out earlier.

But tracing is not necessary for explaining culpable negligence in general, and, as I will argue, it is not necessary for accounting for culpable self-deception. For example, consider that people are subject to general epistemic norms to pay attention to evidence on all sides of a question and even to seek out certain kinds of evidence, at least when the stakes are significant.[35] We have an obligation to take due care in our approach to evidence on important matters. One's failure to do so might then be culpable, even when it does not trace back to an earlier moment of conscious decision not to treat the evidence in a certain way. One might be able to respond to the reasons of taking due care, and yet one fails to do so.

Levy (2004) argues that self-deception is not culpable on the basis of this sort of non-tracing grounds. He begins by suggesting that the relevant conditions under which we could be culpable for such epistemic failures in self-deception cases (such as failures to consider evidence on both sides) are cases in which «(1) the subject matter is important...and (2) that we are in some doubt about its truth» (Levy, 2004, p. 305).[36] Having set out these conditions, he then claims that they are rarely (if ever) met in cases of self-deception. He suggests that «concurrent doubts are ruled out almost by definition: effective self-deception seems to *preclude* the concurrent satisfaction of (2). Successful acts of self-deception leave me in no doubt about the proposition concerning which I am self-deceived» (Levy, 2004, p. 307). And, further, there is no reason to think that they have repressed doubts that were present earlier in the process. It is worth noting that this argument

deception. So in the end, it is not clear what the ultimate disagreement is between Levy and his targets here.

[35] Among the many treatments of this sort of obligation are Fitzpatrick 2008 and Hurd & Moore 2011. For an early version, see Clifford 1877.

[36] Levy argues that condition (1) is not sufficient on the grounds sometimes even when a matter is important, we have no obligation to consider evidence on both sides. As an example, he offers that of the Holocaust scholar who has no obligation to consider the arguments of those arguing that the Holocaust was a hoax. This may be true, but only if the scholar has already considered and responded to the general category of reasoning. Thus, I am not convinced that a second condition is needed.

assumes that the product of self-deception is a belief. I agree that it is, but it is important that not all motivationists do. And one reason for rejecting the claim that belief is a product, for some theorists, is precisely self-deceivers' behavior that is claimed to be characteristic evidence of doubt. More importantly for our purposes here, even if the product of self-deception is a belief with significant consequences for behavior and other attitudes, this is perfectly consistent with doubt. For example, I believe that my spouse and I made the right decision about how much to limit exposure to TV for our children in their early years, but I am not without any doubt about it. In fact, I suspect that for many, any of a large number of parenting decisions is a good candidate for belief with doubt. And the more there is at stake, the more significant the obligation to examine our evidence.

Thus, while there is much work to be done in discovering the correct and complete theory of culpable negligence, there is also no obvious general reason for thinking that self-deception will fail to satisfy its conditions, at least some of the time. To support this claim further, let us turn to the comparative project of examining cases of self-deception alongside other kinds of cases of apparently culpable negligence.

Suppose that instead of being motivated by a desire to believe her son innocent of any possible crime, the grandmother described earlier is simply distracted by loud talk radio whenever her grandchildren and son are in her house. Because she is distracted, she doesn't register well-known signs of abuse, or if she registers them, she doesn't spend the mental energy to investigate further. The details could be filled out in different ways, of course, but as described so far this is easily conceivable as a kind of neglect, and a kind of culpable negligence in not pursuing the evidence of something so important. If asked by a friend whether her son abuses his children, she might sincerely answer that he does not. The question before us is whether there is any difference inherent in the kind of process at work that could make the distraction process and consequences something for which the agent is blameworthy and the self-deception process and consequences something for which she is not. The process in each case might be opaque to the grandmother, the evidence available the same, and her general reasoning abilities identical. Still, one might be tempted to think that the process matters because in the case of self-deception, the motivating desire operates in such a powerful way that it is "irresistible", and so exculpatory. It may be that the desire is strong in many cases, but, again, I think we rarely have evidence that

such desires are irresistible. Further, we lack evidence that it is any stronger than the power of distraction, or of any other potential causes such as intellectual laziness. It is true that the presence of the desire means that there is something at stake for the mother; but there might be something at stake even if the desire does not play a *causal* role. If anything, particularly if she recognizes her own desire, she may have extra reason to be on guard against such a self-deceptive process. (She might also know about herself that she is easily distracted in important situations and so also have extra reason to be on guard here, too, of course.) The upshot thus far is that there is nothing that we know about self-deception that would seem *essentially* excusing or mitigating relative to other kinds of erroneous belief formation against the evidence in cases of significant stakes. In fact, in some cases, self-deception might be more blameworthy than in other such cases.

4. Conclusion

The question of whether and when self-deceivers are responsible for their self-deception and its consequences brings together two independent and important, albeit controversial issues, namely, the nature of self-deception and the conditions under which we are responsible. If the motivationist picture is correct, then it brings the more specific, but no less controversial issue of the conditions for culpable negligence into play, as well. In this paper, I have briefly argued for a particular motivationist account, the Desire to Believe account, and then shown how it forms part of a plausible view of when self-deceivers are responsible and preserves the intuitive idea that in at least some high stakes cases, self-deceivers are responsible for their deception and its consequences. I have also argued that the Desire to Believe account, along with motivationism more generally, offers mutual support to one particular way of developing the powerful idea that we are responsible when we are reasons-responsive agents.

REFERENCES

Alexander, L. and Ferzan, K., with Morse, S. (2009). *Crime and Culpability: A Theory of the Criminal Law*. Cambridge: Cambridge University Press.

Audi, R. (1997). Self-Deception vs. Self-Caused Deception: A comment on Professor Mele. [Open Peer Commentary on Mele 1997] *Behavioral and Brain Sciences*, *20*(1), 104.

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

Bermudez, J. (1997). Defending Intentionalist Accounts of Self-Deception. [Open Peer Commentary on Mele 1997] *Behavioral and Brain Sciences*, *20*(1), 107–108.

Clifford, W. (1877/2008). The *Ethics of Belief*. [reprinted in L. Pojman & M. Rea (Eds.) (2008) *Philosophy of Religion: An Anthology*. Boston: Wadsworth]

Davidson, D. (1986). "Deception and Division." In J. Elster (Ed.), *The Multiple Self*. Cambridge: Cambridge University Press, 79–92.

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

Deweese-Boyd, I. (2010). *Self-Deception*. The Stanford Encyclopedia of Philosophy (Fall 2010 Edition). <http://plato.stanford.edu/archives/fall2010/entries/self-deception/>.

Fernandez, J. (2011). Self-Deception and Self-Knowledge. *Philosophical Studies*.

Fischer, J.M., & Ravizza, M. (2000). Responsibility and Control: A Theory of Moral Responsibility. Cambridge: Cambridge University Press.

Fingarette, H. (1969). *Self-Deception*. Berkeley: University of California Press; reprinted, 2000.

Fitzpatrick, W. (2008). Moral Responsibiilty and Normative Ignorance: Answering a New Skeptical Challenge. *Ethics*, *118*(4), 589–614.

Funkhouser, E. (2005). Do the Self-Deceived Get What They Want?. *Pacific Philosophical Quarterly*, *86*(3), 295–312.

Gendler, T.S. (2007). Self-Deception as Pretense. *Philosophical Perspectives*, *21*(1), 231–258.

Jenni, K. (2003). Vices of Inattention. *Journal of Applied* Philosophy, *20*(3), 279–295.

Johnston, M. (1988). Self-Deception and the Nature of Mind. In B. McLaughlin & A.O. Rorty (Eds.), *Perspective on Self-Deception*. Berkeley: University of California, 63–91.

Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Levy, N. (2004). Self-Deception and Moral Responsibility. *Ratio* (new series), *17*(3), 294–311.

Mele, A.R. (1987). *Irrationality: An Essay on Akrasia, Self-Deception, and Self-Control*. Oxford: Oxford University Press.

Mele, A.R. (1997). "Real Self-Deception" and "Author's Response". *Behavioral and Brain Sciences*, *20*(1), 91–102, 127–136.

Mele, A.R. (1999). Twisted Self-Deception. *Philosophical Psychology* 12: 117–137.

Mele, A.R. (2000). Self-Deception and Emotion. *Consciousness and Emotion, 1*(1), 115-137.

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

Mele, A.R. (2009). Have I Unmasked Self-Deception or Am I Self Deceived?. In C. Martin (Ed.), *The Philosophy of Deception*. Oxford: Oxford University Press, 260–276.

Moore, M. & Hurd, H. (2011). Punishing the Awkward, the Stupid, the Weak, and the Selfish: The Culpability of Negligence. *Criminal Law and Philosophy*, *5*(2), 147–198.

Nelkin, D. (2002). Self-Deception, Motivation, and the Desire to Believe. *Pacific Philosophical Quarterly*, *83 (4)*, 384–406.

O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. Oxford: Oxford University Press.

Pears, D. (1984). *Motivated Irrationality*. Oxford: Oxford University Press.

Pereboom, D. (2001). *Living Without Free Will*. Cambridge: Cambridge University Press.

Scott-Kakures, D. (2002). At Permanent Risk: Reasoning and Self-Knowledge in Self-Deception. *Philosophy and Phenomenological Research*, *65*(3), 576–603.

Scott-Kakures, D. (1996). Self-Deception and Internal Irrationality. *Philosophy and Phenomenological Research*, *56*(1), 31–56.

Sharpsteen, D. & Kirkpatrick, L. (1997). Romantic Jealousy and Adult Romantic Attachment. *Journal of Personality and Social Psychology*, *72* (3), 627–640.

Sher, G. (2009). Who *Knew? Responsibility Without Awareness*. Oxford: Oxford University Press.

Wallace, R.J. (1994). *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.