(This is the postprint of a review published in *Philosophy*, vol. 95(1), in 2020)

Self-Deception
By Eric Funkhouser. Routledge, 2019. (From the New Problems of Philosophy series, edited by José Luis Bermúdez.)
ISBN: 978-1138506121

This new book on self-deception by Eric Funkhouser is not a treatise but an introduction to and survey of the field of study. As such, it is the first of its kind and the first book of any sort to be published on self-deception for quite some time. It will be particularly useful for people new to the field, partly for the good job it does of bringing presentable order to the variety of accounts of self-deception that have proliferated in recent decades. But it is also well worth reading for advanced researchers for its very detailed, observant and balanced commentary on a variety of theories and for its many sharp criticisms.

Funkhouser begins by setting out the problems, puzzles and connections to other topics that motivate interest in self-deception, and by detailing common putative self-deceptive biases. He identifies the 'Basic Problem' of self-deception as being 'to show how active measures to deceive can be taken against a single self with actual success in acquiring the deceptive belief' (29). But it turns out that this is a pseudo-problem, since all 'solutions' to it work by rejecting or at least seriously qualifying one or another of its presuppositions. Funkhouser divides these into three categories, approaches that *sacrifice deception* (intentional or purposive manipulation), that *sacrifice the self*, or that *sacrifice belief* (29-30). According to the first, self-deceptive beliefs are caused mostly by sub-intentional motivational processes. The second retains the idea that the belief is intentionally brought about and appeals to some sort of 'divided mind' thesis to explain how this is possible. And the third relinquishes the assumption that self-deceptive processes result in a false or unjustified belief.

These different theoretical approaches offer us a variety of different pictures of what self-deception is, but if I understand him correctly, Funkhouser advocates an inclusive view according to which they all represent different kinds of self-deception, though some count as deceptive to a greater or 'more robust' degree than others (61, 137), with the phenomena described in deflationist accounts being at the lower end of the spectrum of deceptiveness. We also get a chapter on responsibility for self-deception and one on the benefits and harms of self-deception. These are some of the best discussions of these topics out there, and they introduce some useful new concepts like baited self-deception and enabled self-deception.

The only criticism I would make of the book concerns Funkhouser's way of classifying and distinguishing the different broad theoretical approaches to self-deception. The above tripartite distinction is indeed apt, and we are led to expect that it maps onto the distinctions between deflationist, intentionalist, and revisionist theories of self-deception discussed in chapters three, four and five respectively. However, some of the theories discussed in the revisionist chapter hold that self-deception results in false or unjustified belief, such as the theories of D. L. Smith, and von Hippel and Trivers. Instead, these are said to be revisionist concerning the deception element. That makes it surprising that they were not placed in the chapter on deflationism, which 'sacrifices deception'. Further, Mark

Johnston's theory is discussed under deflationism and D. L. Smith's under revisionism, though these views are very similar; Smith says that Johnston's view is closest to his own.[1]

This is just an organisational matter, but a more philosophically significant issue arises with how Funkhouser understands the distinction between deflationism and intentionalism. These approaches share the assumption that self-deception results in false or unjustified belief, but they differ in how they see that belief as being produced. In chapter two however, 'The basic problem and a conceptual map', Funkhouser contrasts intentionalism with motivationalism, but this seems to be just another word for 'deflationism'. At least he doesn't explain what the difference is between them if there is one, and the terms are commonly used interchangeably in the literature. Further, he states that motivationalism is equivalent to anti-intentionalism (55). *Anti*-intentionalism, moreover, just means 'against intentionalism': it is the denial of the intentionalist's thesis that the false/unjustified belief is intentionally produced (or sustained) by the subject. So by the transitivity of identity that is what deflationism is.

However, Funkhouser also says that in explaining the irrational belief, deflationism 'abandon[s] strong agential involvement' (85) or is 'on the low end of the spectrum when it comes to agential involvement' (86). The idea here is that 'desires and emotions can … act as mental causes that produce effects through associations or simple brute causal connections' (59). This contrasts with how desires 'generate actions by presenting us with goals' (58), as they do in the intentionalist's picture, where the self-deceiver has the goal of having the self-deceptive belief.

But this is an entirely different and narrower idea from the denial of intentionalism. It is quite possible to deny that self-deceivers intentionally deceive themselves while also holding that their unjustified beliefs are due to their intentional actions, so long as the guiding intention isn't represented to be an intention to deceive. In other words, we can distinguish two kinds of deflationism or non-intentionalism, an agentive and a non-agentive kind.[2] An agentive form of deflationism is something that gets omitted from Funkhouser's listing of the available explanatory options.

This ambiguity then infects how Funkhouser describes deflationism's opposite, intentionalism, when he writes, '*Intentionalism*: Self-deception is [or better, involves] an intentional action. The self-deceived have a specific intention to deceive themselves or to acquire the false or unsupported belief' (55). But this description runs together two different ideas. The idea that self-deceptive beliefs are produced by intentional actions is different from the idea that they are produced by deceptive intentions, and one can accept the former while rejecting the latter (but not vice versa).

Funkhouser usefully distinguishes between world-directed motives and mind-directed motives (66). The first is a desire for the world to be a certain way, and the second is a desire for one's mind to be a certain way, and different accounts of self-deception put explanatory emphasis on one or the other. One reason why Funkhouser doesn't take seriously the possibility of an agentive kind of deflationism is that deflationists like Annette

[1] David Livingstone Smith, 'Self-Deception: A Teleofunctional Approach', *Philosophia* **42** (2014), 182, note 2.

[2] See Kevin Lynch, 'An Agentive Non-Intentionalist Theory of Self-Deception', *Canadian Journal of Philosophy* **47** (2017).

Barnes and Alfred Mele have typically explained with world-directed motives–the desire that *p* rather than the desire to believe that *p*–and Funkhouser doesn't think that the desire that *p* can give the subject any relevant goal. As he says, 'for world-directed motivationalists the acquisition of a belief that [*p*] might terminate the self-deceptive process, but it is hard to think of this as goal satisfaction if the motivation was a desire that the world be such that [*p*]. Believing does not make it so!' (72).

Of course, desiring that *p* often gives rise to the goal of ensuring that *p*. But this doesn't happen in self-deception, where matters are typically out of one's hands. However, there is another way the desire that *p* can give rise to a goal that could cause the self-deceptive belief. When one strongly desires that *p* (where it's beyond one's control whether *p*) and that proposition is thrown into doubt, so that one is anxiously wondering whether *p*, it is then quite natural to want there to be evidence or reasons that would confirm that *p*, since it would be a relief to have *p* confirmed. This can give rise to the goal of finding *p*-confirming evidence, so as to assure oneself or put one's mind at ease that *p*, leading to a biased evidence search. In this manner, we can see self-deceptive biases as being produced by intentional behavior that is not driven by a mind-directed motive.

The failure to distinguish agentive from non-agentive forms of deflationism then leads to a one-tracked interpretation of the various psychological processes that deflationists mention as explaining self-deceptive beliefs. One of these discussed by Funkhouser is memory biases or selective recall (10-11), where one mostly remembers what is conducive to believing one thing rather than another. One way to interpret this is as being due to some sub-personal and sub-intentional mechanism, where the desire that *p* somehow makes memories conducive to believing that *p* more accessible, so that they just come to mind more easily when one thinks of the matter. Funkhouser, who says that memory biases are 'harder to envision as intentional' (146), is probably thinking along these lines. But an alternative view of what is going on, which is also deflationist, is that the subject is *searching her memory specifically* for data supportive of what she wants to be true, *in the hope of confirming it*. And this searching would be an intentional, goal-directed activity, directed towards the goal of finding any *p*-confirming evidence. With this interpretation of memory biases we would be on the high end of the spectrum of agential involvement and would not be dealing with brute causal connections.

In fairness to Funkhouser, this problem with classifying different approaches to explaining self-deception is not of his own making. It is standard practice in the literature to conflate the intentionalism/anti-intentionalism distinction with the agency/anti-agency view distinction, and to set up the debate on how self-deceptive beliefs arise as being between intentionalism and non-agentive deflationism, ignoring the possibility of an agentive deflationism.

This is maybe the only lacuna in what is otherwise a very thorough survey. It is a book peppered with insightful commentary and interesting ideas (such as the intriguing suggestion that we might be able to determine the degree of agentive involvement in putative cases of self-deception by seeing how the subjects' biases are affected by cognitive load tasks (108-109)). It deserves to become the key reference work in this topic.

Kevin Lynch