INTENTIONALISM AS A THEORY OF SELF-DECEPTION

Ivan Cerovac
University of Trieste, Department of Philosophy, Trieste, Italy
e-mail: ivan.cerovac@phd.units.it

Abstract:
Is self-deception something that just happens to us, or is it an intentional action of an agent? This paper discusses intentionalism, a theory claiming that self-deception is intentional behavior that aims to produce a belief that the agent does not share. The agent is motivated by his belief that \( p \) (e.g. he is bald) and his desire that \( \neg p \) (e.g. not to be bald), and if self-deceiving is successful, the agent will end up believing \( \neg p \). Opponents of intentionalism raise two different objections: it seems that self-deceiver should then simultaneously hold two incompatible beliefs (namely, that \( p \) and \( \neg p \)), as well as simultaneously intend the deception and be unaware of it. This paper reviews possible answers to anti-intentionalist objections (temporal partitioning, psychological partitioning, and the attentional strategy account) and offers guidelines to strengthen intentionalist claims.

Key words: Self-deception, Intentionalism, Anti-Intentionalism, Mind partitioning, Attentional strategy.

I’m not fat. I’m big-boned.  
– Eric Cartman

Eric Cartman, a fictional character from South Park series, believes that he is not fat, despite the overwhelming available evidence to the contrary (all his friends and teachers think he is fat, and his weight greatly exceeds the one recommended for his age and height). Moreover, Eric has come with what he holds to be a reasonable explanation for their mistake—he is not fat, but big-boned. We should also consider the following: (i) his epistemic standards are not defective in any relevant way—given the described evidence he should be able to conclude that he is fat, and (ii) he has a strong desire not to be fat. Is Eric self-deceiving himself (or being self-deceived)? To answer this question, we must first define self-deception. But how are we to go about it?

Alfred Mele distinguished two main approaches to the analysis of self-deception: the lexical and the empirical (or example-based) approach. While the former—usually associated with traditionalists—“starts by establishing a definition of deception from the interpersonal case, and uses it to deduce the meaning of self-deception” independent of the usage of expression by ordinary people, the latter—usually associated with deflationists—“starts by scrutinizing representative examples of self-deception and attempts to identify their essential common features.” Both positions generally agree that self-deception is “the acquisition and maintenance of a belief (or the avowal of that belief) (i) in the face of strong evidence to the contrary (ii) motivated by desires or emotions favoring the acquisition and retention of that belief.” It seems that Eric is indeed deceiving himself (or being self-deceived), since he is believing something against his better judgment because of his strong desire.

But why does he end up believing that \( \neg p \) despite the fact that he has both adequate cognitive capabilities and relevant evidence to conclude that \( p \)? Certainly there has to be some reason why Eric

---

1 Ariela Lazar (1999) distinguishes these two formulations and connects them with two main theories of self-deception: “deceiving oneself” with intentionalism and “being self-deceived” with anti-intentionalism.
ends up falsely believing that he is not fat. First of all, we must perceive that the phrase “Eric has a reason to believe that he is not fat” has more than one meaning: “[a] reason of the first sort is evaluative: it provides a motive for acting in such a way as to promote having a belief, and a reason of the second kind is cognitive: it consists in evidence one has for the truth of a proposition.” In many examples of self-deception, including the former one, irrationality arises when someone has only an evaluative (motivating) reason for a belief that \( p \), and he acquires that belief despite the fact that he has no cognitive (normative) reason to believe it. The problem with self-deception emerges when a person starts having thoughts that emphasize the obtaining of cognitive (normative) reasons for a belief because of desires or motivating reasons for holding it. Self-deception is the process (or the state) that leads from a desire or a motivating reason to a reason that the self-deceiver regards as normative for a certain belief. It usually consists of intense epistemic activity that attempts to justify the false belief that \( \neg p \)—since initially the self-deceiver has better (epistemic) reasons to believe that \( p \), he has to thoroughly reconsider the evidence he has (find new evidence) in order to end up believing that \( \neg p \). There are many strategies one can use to manipulate the belief-formation process to favor a certain belief, such as selective evidence gathering and selective focusing of attention. I will not discuss these strategies in detail; we should rather inquire into how one applies them. Is Eric intentionally gathering evidence selectively, or is this cognitive bias something that happens to him? Is Eric the author or the victim of self-deception?

Contemporary discussion of self-deception distinguishes two central positions—intentionalism and anti-intentionalism. Intentionalists compare self-deception with interpersonal deception, intentional behavior that aims to produce a belief that the agent does not share. Unlike interpersonal deception, where A intentionally gets B to believe that \( p \) while A knows that \( \neg p \), in self-deception A is simultaneously the deceiver and the deceived—he intentionally gets himself to believe that \( p \) while also knowing or believing that \( \neg p \). Anti-intentionalists, on the other hand, consider self-deception to be a form of motivationally biased belief, not at all structurally similar to intentional interpersonal deception. They hold that our false beliefs are not the result of an intentional process, but unintentionally motivated by a desire or emotion.

This paper is divided into three distinct sections. The first provides a summary of main arguments that support intentionalism as a theory of self-deception. Following Davidson (1985) and Bermudez (2000), I first defend intentionalism through inference to the best explanation using the lexical approach to the analysis of self-deception, and then, noticing the paradigmatically selective nature of self-deception, argue that only intentionalist approaches can explain why self-deception took place in that particular situation. In the second section I consider the most common objections to intentionalist approach—the “static” and “dynamic” (or “strategic”) paradoxes. Since there is more than one way to answer these objections, I briefly sketch possible intentionalist accounts that avoid these paradoxes: Bermudez’s (2000) “temporal partitioning,” Davidson’s (1982) “psychological partitioning,” and finally the “attentional strategy.” In the final section, I shall briefly summarize the main explanatory virtues of intentionalism and provide some additional arguments that support this claim.

1. The advantages of the intentionalist approach

Considering self-deception analogous to interpersonal deception enables us to distinguish it from a mere error, since in self-deception acquisition and maintenance of a false belief is intentional. Self-deception can be easily shown by distinguishing between belief—a non-intentional but causally produced mental state that one cannot be held responsible for—and acceptance—an intentional process based on reasons and subject to one’s control. Self-deception would thus be a process in which a person acquires and maintains a false belief (i) despite many reasons available to her, and (ii) tries to accept false belief by means of filtering the information available in various ways. The intentionalist approach gives us a clear distinction between self-deception and other sorts of motivated believing, like wishful thinking and weakness of warrant. Davidson thus distinguishes wishful thinking—the

---


7 For a detailed distinction between self-deception and other sorts of motivated believing, see Davidson's paper “Paradoxes of Irrationality.”
sequence in which the desire that $p$ produces the belief that $p$—and self-deception, which requires "the agent to do something with the aim of changing his own views." The self-deception must not only be self-induced, it also must be intentional.

It is not self-deception simply to do something intentionally with the consequence that one is self-deceived, for then a person would be self-deceived if he read and believed a false report in a newspaper. The self-deceiver must intend the "deception." Davidson explains how self-deception occurs.

A has evidence on the basis of which he believes that $p$ is more apt to be true than its negation; the thought that $p$, or the thought that he ought rationally to believe that $p$ motivates A to act in such a way as to cause himself to believe the negation of $p$. \[\ldots\] All that self-deception demands of the action is that the motive originates in a belief that $p$ is true (or that the evidence makes it more likely to be true than not), and that the action be done with the intention of producing a belief in the negation of $p$. Finally, and it is especially this that makes self-deception a problem, the state that motivates self-deception and the state it produces coexist; in the strongest case, the belief that $p$ not only causes a belief in the negation that $p$, but also sustains it.

Although most philosophers who discuss the self-deception problem here include Davidson's example of Carlos and his driving license test, I believe another example, also made by Davidson, is more appropriate: the bald man example. It displays not only how a desire, along with an appropriate belief, can induce self-deception and lead to an agent believing something that is false, but also how this process can continue indefinitely into the future, with the belief that $p$ constantly sustaining the belief that $\neg p$. D believes that he is bald and has good reason to believe that he is bald. Since he does not want to be bald (and does not want to believe that he is bald) he chooses views and lighting that create a hirsute appearance. Now, with his new hairstyle, D starts to believe that he is not all that bald. His belief that he is bald, along with his desire not to be bald, have led to selective evidence gathering, and as a result D ends up believing that he is not bald. However, he keeps choosing views and lighting that create a hirsute appearance. Why does he keep choosing them if he now truly and completely believes that he is not bald? The answer seems rather obvious: though D now believes that he is not bald, he still has a belief that he is, and it is exactly this belief—along with his desire not to be bald—that motivates him to keep choosing certain views and lighting. It seems that anti-intentionalists cannot explain self-deception in this example by the usage of desires and emotions alone; it is certainly correct that our desires play a great role in self-deception, but it also seems that they alone are not enough. Suppose D had the desire not to be bald from his childhood, for a very long period of time. 10 years ago, when he did not have good reasons to believe that he is bald (probably because he wasn't) he still had the desire not to be bald, but it did not motivate him to choose views and lighting that create a hirsute appearance.

Most intentionalists claim that D's self-deception is the result of his practical reasoning. If asked why D intentionally forms a belief that is undermined by the weight of the evidence, they suggest that the irrational belief is acquired in order to attain a goal that is frustrated by the presence of the rational belief. Thus Davidson introduces the Carlos example in order to elaborate how a belief (or the perception that one has sufficient reasons for a belief) can sustain (cause) a contrary belief. Davidson writes

So he has a perfectly natural motive for believing he will not fail the test, that is, he has a reason for making it a case that he is a person who believes he will (probably) pass the test. His practical reasoning is straightforward. Other things being equal, it is better to avoid pain; believing he will fail the exam is painful; therefore (other things being equal) it is better to avoid believing he will fail the exam.\[^{10}\]

Self-deception is also paradigmatically selective. "Any explanation of a given instance of self-deception will need to explain why motivational bias occurred in that particular situation."\[^{11}\] According to

---


9 Ibid.

10 Ibid.

the anti-intentionalist account, a strong desire and a low acceptance/high rejection threshold for the hypothesis that \( p \) are enough to induce self-deception and make a subject believe that \( p \) (no intention from the subject is required). But there are also many cases in which we have both the desire and a low acceptance/high rejection threshold, but we don’t end up believing that \( p \). Why does a person end up self-deceiving in one example, and not in other, while she had an equally strong desire and low acceptance/high rejection threshold for both? Bermudez (2000) concludes that “it (also) requires an intention on the part of the self-deceiver to bring it about that he acquires the belief that \( p \).” Since anti-intentionalists cannot solve the selectivity problem, this solution offered by intentionalists proves to be a strong argument for their position.

Since (unlike intentionalism) anti-intentionalism is often argued for indirectly through attacking incoherencies in the intentionalist account, in the next section we shall review these objections and possible solutions.

2. Common objections to the intentionalist approach

In this section we shall discuss the most common objections to intentionalist approach—the “static” and “dynamic” (or “strategic”) paradoxes. Both paradoxes arise from the analogy of self-deception and interpersonal deception. Davidson (1985) claims that “the only intentions a liar must have are these: (1) he must intend to represent himself as believing what he does not (for example, and typically, by asserting what he does not believe), and (2) he must intend to keep his intention hidden from his hearer.” In other words, (1) the deceiver must believe the negation of the belief he wants to induce in the hearer, and (2) he must intend to deceive the hearer, and the hearer must not know about the deceiver’s intentions. Since in self-deception both the deceiver and the hearer are the same person, this means that (1) the self-deceiver must believe both the belief he wants to induce in himself and the negation of that belief (he must believe both \( p \) and \( \neg p \)), and that (2) he must intend to deceive himself and simultaneously be ignorant about his own intentions.

The first requirement leads to the “static” paradox—it seems impossible to consciously believe both \( p \) and \( \neg p \) at the same time. The self-deceiver must (as a deceiver) believe that \( \neg p \), while at the same time (as a hearer) believe that \( p \). This, however, seems to be an impossible state of mind, because it violates the principle of non-contradiction.

The second requirement leads to the “dynamic” paradox—it seems impossible to simultaneously intend the deception and remain unaware of our own intentions. Thus, even if intentionalists somehow manage to avoid the “static” paradox by asserting that the self-deceiver gets a change of mind (believes that \( p \), but after reconsideration of the evidence ends up believing that \( \neg p \)), so he never simultaneously believes both \( p \) and \( \neg p \), they still have to face the fact that the intention to deceive ourselves renders the process self-defeating.

There are several intentionalists accounts that avoid these paradoxes, and we shall briefly review three main responses.

Temporal Partitioning

One of the possible intentionalist strategies is to admit that it is impossible to simultaneously believe both \( p \) and \( \neg p \), but claim that this does not imply that self-deception is impossible, since it is often extended over time. Thus a self-deceiver can believe that \( p \) in \( t_1 \) and that \( \neg p \) in \( t_2 \) without ever simultaneously believing both \( p \) and \( \neg p \) and falling into the ‘static’ paradox. Though the “dynamic” paradox cannot be avoided so easily, Bermudez claims that “one can simply lose touch with an intention while one is in a process of implementing it, particularly when that implementation is a long drawn out process.” And even if one still remembers his own intention to deceive oneself, he might still be so overwhelmed with new evidence he had acquired through self-deception that he no longer cares about his former intention.

Psychological (Mind) Partitioning

Another possible intentionalist strategy is to directly confront the ‘static’ paradox and claim that it is possible to simultaneously believe both \( p \) and \( \neg p \) without ever making the conjunction \( (p \land \neg p) \). Davidson thus claims that the mind can be partitioned into quasi-independent structures, and that there can

---


be non-logical causal relations between them.\textsuperscript{14} We can have two contrary beliefs, but placed in different “brain areas”; this is Davidson’s solution to the “static” paradox. The “dynamic” paradox proves to be a serious problem for the intentionalists; it seems impossible to deceive oneself since it would require doing something with the intention that the interlocutor does not recognize this very intention. However, Davidson claims that there is a distinction between self-deception and lying to oneself: while the latter entails the existence of a self-defeating intention (and is thus struck by the “dynamic” paradox), the former “pits intention and desire against belief, and belief against belief.”\textsuperscript{15} Therefore the subject of self-deception tries to find other reasons to justify his selective acquisition of the available information; he is not lying to himself that he is bald (because he acquires reasons to believe that he is not bald after he has chosen views and lighting that create a hirsute appearance), and he chooses views and lighting because of some other selectively chosen evidence (e.g. someone once told him that he looks better that way). The deceptive subsystem is thus hidden from consciousness in a complex network of motivated and interconnected reasons, and the “dynamic” paradox is avoided.

The Attentional Strategy

Yet another possible intentionalist strategy includes intentional manipulation of our attention, as presented by Lynch (2009). “A person who believes or at least suspects that \( p \), but who wishes to believe that \( \neg p \), by turning his attention away from the unwelcome considerations supportive of \( p \), and by attending to welcome considerations supportive of the contrary \( \neg p \), may end up losing the (conscious) belief that \( p \) and acquiring the belief that \( \neg p \).”\textsuperscript{16} Consider a student who has good reasons to believe that he should study for his final exam, but does not want to do so. If he tries to convince himself that he will probably pass the exam without studying he faces the “dynamic” paradox; however, if he turns his attention to something else, like playing football or watching movies, he successfully avoids it. When he is playing football, he is both successfully self-deceived and unaware of the deception he has just made. “Although his distraction seeking can be done intentionally, this does not imply that agent is intentionally not thinking about the unwelcome evidence when she succeeds in distracting herself from thinking about it.”\textsuperscript{17}

3. Conclusion

This paper emphasizes two main arguments for the endorsement of intentionalism: its attitude towards the nature of self-deception and the development of an intuitive explanation of that phenomena and supports our intuitive idea that the self-deceiver should be responsible for his self-deception, and the inclusion of intention offers a solution to the problem of the paradigmatically selective nature of self-deception. In the second section we have considered possible answers to the “static” and “dynamic” paradoxes. Although some of these answers may raise objections, the discussion is still active and there is no final conclusion. Another suggestion is presented by Bach (1981), who claims that the self-deceiver does not intend the deception, but nonetheless acts intentionally. “Although the self-deceiver does what he does intentionally, he does not do it under the description of ‘deceiving myself’ or anything of that sort. Rather, he is motivated to avoid the thought that \( p \) but is unaware of (or denies the impact of) this motivation and of his uncharacteristic violation of his own rational standards.”\textsuperscript{18} This suggestion successfully avoids the two paradoxes and is somewhat similar to Davidson’s position, especially in respect to unconscious beliefs and intentions. I believe that psychology has yet to provide its main contribution to this discussion, and that strategies involving unconscious beliefs and desires have significant potential.

References


\textsuperscript{17} Ibid.


