Does self-deception constitute a threat to the distinguishing kind of rationality human beings enjoy? I propose a compromise between the deflationary solutions to the alleged puzzles self-deception has long been taken to give rise to and some virtues of the competing account -namely, the intentionalist view. In order to fulfil the task, I argue as to precisely what is left to blame in self-deception once we have made sense of why intentionalism fails to capture the nature of the phenomenon.

Keywords: self-deception, belief, irrationality, rationality.

The trick of desire – which avails itself of any irrelevant scepticism, finding larger room for itself in all uncertainty about effects, in every obscurity that looks like the absence of law.
Who could say that the death of Raffles had been hastened? Who knew what would have saved him?
George Eliot, Middlemarch

1. Introductory overview

A husband believes that his wife is not having an extra-marital affair, as she has in fact been having, despite the evidence at his disposal could easily alert him and make him form the belief that she is unfaithful.¹ A skilled oncologist believes that she has not a cancer, despite the fact that she displays symptoms that she is normally able to judge as diagnostic of

---

¹ This is one of the most cited examples throughout the literature. For a survey of similar examples, see A. Lazar, Deceiving Oneself or Self-Deceived? On the Formation of Belief 'Under the Influence', «Mind», 108, 1999, pp. 285-90.
cancer in her patients. A mother believes that her son is innocent of a crime for which he has been convicted after a full confession. And again, a girl believes that she is displeased with a man with whom she is instead in love and she believes this despite the fact that in the circumstance in question she does and feels whatever she usually does and feels in front of a boy she likes. A wife methodically attends to the death of her alcoholic husband by letting him drink excessively, but she now believes that she did nothing decisive to hasten his death.

All the cases sketched above are normally taken to be representative of self-deception. They show people believing propositions which are not only false, but also extremely weakly warranted by the believers’ own lights, given the evidence, or clues of easily accessible evidence, they themselves cannot - so we judge - have missed or overlooked. Another clear feature of the propositions they come to believe is the desirability of their contents. By hypothesis, the believers in question are sufficiently competent epistemic agents and in many other circumstances they repeatedly show and have shown to be perfectly able to seek, evaluate and use similar, relevant evidence to form beliefs that are justified in light of it. Their epistemic standards are not defective in any relevant way. That is, their conceptual capacities are integral, their sensitivity to reasons efficient. Thus, if they really falsely believe what they ended up believing, there must be a reason (or at least some sort of intelligible cause) why their usual epistemic standards, such as


3 A similar example can be found in D. H. Sanford, Self-Deception as Rationalization, in Perspectives on Self-Deception, ed. by B. McLaughlin – A. O. Rorty, cit., pp. 156-169; similar examples are also discussed in A. Mele, Self-Deception Unmasked, Princeton University Press, Princeton 2001; and in J. V. Canfield – D. F. Gustafson, Self-Deception, «Analysis», 23, 1962, pp. 32-36.


6 The possibility of being self-deceived in believing a proposition which happens to be true will not be discussed in this essay.

7 They could only be insincere. Deciding whether a believer is only mendacious or self-deceived is an interesting task, but it will not be tackled here.
the requirement of total evidence for inductive reasoning, are – so it seems - inoperative or suspended. In general, we need an explanation as to why similar doxastic outcomes can be brought about – one which be both sufficiently realistic and informative.

I assume that we as interpreters wish to find out in their overall psychological predicament one or some internal reasons, that is to say, reasons they as epistemic agents recognize as such and have employed in their reasoning towards the doxastic conclusion they reached. This discovery would be paramount to our understanding of them and to preserving the minimal rationality we credit each other with, abnormal cases apart. At the same time, we also expect that these internal reasons must be somehow defective or distorted as not to count as normative reasons, if the irrational outcome resulted. It is false that ‘tout comprendre est tout pardonner’. A motivating reason can have a causal effect on our thinking and behaviour and thus be explanatory of that very thinking and behaviour, and still not be a normative reason to think and act as we do.

The thesis hinted at above according to which the usual epistemic standards are not applied in self-deception is only one of the possible working hypotheses a theorist may begin with when analysing self-deception. Part of my attempt will be in fact to vindicate the idea that self-deception requires far more than a ‘mere’, ‘immediate’, ‘non agential’ suspension of usual epistemic activity; rather, it is my view that self-deception consists of an intense and often psychologically wearing epistemic activity. If there is at some point of the self-deceptive process a suspension of them at all, this must nonetheless be sustained by a host of ‘justifying reasons’ that help the self-deceivers wrongly believe that ‘normal standards are not to be applied in the situation they are involved in’. This is the ‘crucial’ self-deceptive belief, more than the false belief they come to acquire. For self-deceivers’ rationality requires that they give themselves an answer to the question as to why the correct doxastic conclusion ‘cannot’ be achieved in the specific case they confront. This distinguishes self-deception from other more immediate forms of motivated irrationality, as we will see. The self-deceiver’s constraints over

\[8\] See D. Davidson, *Incoherence and Irrationality*, in «Dialectica», 39 (4), 1985, pp. 345-354. The argument is that by a priori assuming the rationality of human beings, we naturally look for rationalizing explanations of their behaviour.


\[10\] In general, to say that one action of belief is rationalized by a series of considerations does not equal to say that the action or the belief are rational.
the sufficiency of this ‘covering beliefs’ concerning their epistemic judgments themselves may be more or less demanding, more or less articulated, depending on the conceptual sophistication, the actual circumstances and the rationalising satisfaction threshold of the believer. When explaining self-deception, we must be prepared to allow for the possibility of its occurring in different degrees. In any case, if there is no justifying reason about the suitability of the suspension of the usual epistemic standards we may well be in front of a form a motivated irrationality, though not yet in front of an instance of full-blown self-deception.

Contrary to many positions in the literature, I take this very epistemic endeavour that we call self-deception to be largely conscious. Self-deception is paradigmatically achieved by means of a tentacular web of rationalizing covering stories to which the self-deceiver carefully and obstinately commits his reasoning activity. Otherwise, again, it deserves the name of some other less specific, though common and real, motivated irrationality. Many of the above examples can be hard-pressed enough to show that the self-deceiver’s participation to his self-deception is active and that self-deception is not entirely accomplished, so to say, ‘behind one’s back’, nor is it achieved solely – as I will say - through the effect of motivation over cognition.

Still, it is true that in self-deception ‘something’ escaped the self-deceiver’s control. Among those who go for an ‘active’ characterization of self-deception, it is typically maintained that self-deception must involve some sort of unawareness: some have argued for the unawareness of the belief that would be correct to form and that self-deceivers indeed have formed, being the formation of it just the triggering cause of the self-deceptive process’ starting-up.11 Others have argued for the unawareness of the entire process.12 How to explain it may involve deploying a number of both static and dynamic hypotheses, respectively both on the architecture and the working of the mind. Their rationale in the theoretical economy of these theories and how my account departs from them will be clarified in section 2 and later in section 4 and 5.


On my view, instead, the failure of self-knowledge that self-deception involves, which is distinctive of it and on which self-deception essentially runs, does not lie either in the unconscious character of the starting cognitive state (ranging from a mere suspect to a belief) or in the unawareness of the whole cognitive process - as it happens in cases of mere motivationally biased beliefs - but elsewhere. It is this failure that I take to be what we blame in self-deception; indeed, what the self-deceiver himself will blame later, when he comes to realize to have undertaken a self-deceptive reasoning. As already mentioned, I will count on a distinction between self-deception and motivationally biased beliefs: self-deception appears as a sub-set of the latter, major class. The distinction will be central in order to clarify that self-deception is best understood as a process, and that, if successful, it generates motivationally biased beliefs whose irrational status is on a par with those produced by means of other biased psychological processes triggered by a motivation, such as fanaticism, partiality and ‘hot’ biases in general, while the kind of process which leads the believer to them makes up for the difference – a difference reflected by what we specifically blame in either case.

If self-deception is correctly describable as a largely conscious reasoning process, and in general if there is anything to blame at all in self-deception, the question as to whether the self-deceivers are the ‘victims’ or the ‘authors’ of self-deception becomes more pressing. Depending on the answer, we get an insight into the conception of human rationality and cognition we are willing to subscribe to. The polemic field contending the correct theory of self-deception thus far has been divided into two main groups: ‘intentionalists’ (or ‘traditionalists’) and ‘anti-intentionalists’ (or ‘deflationists’). Intentionalists deem self-deception to be a psychological process which takes place by means of the intentional activity of the believer which ends up in a doxastic state toward which the believer had extremely good reasons not to end up: at the very least, the minimal assumption of an intentionalist is that the self-deceiver plays a decisive role in believing that p, where p is the self-deceptive proposition, by ‘intentionally leading himself to get convinced’ of its truth in spite of easily accessible or directly

---


14 The term ‘self-deception’ is often uncritically used both for the process and for the final doxastic state. I will try to clarify from time to time whether I refer to the former or to the latter.

15 A bias is generally called ‘hot’ when accompanied or triggered by an emotional state of some kind.
available evidence suggestive of its falsity. There is room for much stronger claims within the intentionalist field, e.g. that the self-deceiver believes what he ‘knew’ it was not so; or that the self-deceiver continues to believe the falsity of the self-deceptive proposition while believing its truth as well – he still ‘knows’ it; or, even, that the self-deceiver believes its truth ‘just because’ he never ceases to believe its falsity. In that case, it is assumed that the self-deceiver «must intend his deception».16 I will explain in section 2 how these further theses are supposed to work and why they have been advanced.

Anti-intentionalists, on the other hand, deny that such an intention is either necessary or sufficient, or both, to explain self-deception and also deny that it is phenomenologically present in the self-deceiver’s psychology, either consciously or unconsciously. Their denials are of course shaped by the specific intentionalist claim that their proponents are mainly willing to reject, but much of this in section 3. In general, however, a core claim they deny is that a «desire to believe» a certain proposition \( p \) must be consciously held as the ‘reason’ on which the self-deceiver would intentionally act, instead of as the unintentional ‘cause’ that triggers the process. I will explain in full detail what deflationists object to intentionalists in section 3.

What briefly introduced up to now already makes us have in sight the crucial problem the debate generates. It is tightly linked up to human rationality and it is, roughly, the following: if the intentionalists are right, human rationality seems to be deeply threatened: as believing and intending are involved in self-deception, how can we make sense of the possibility of intentionally leading us to believe what we must at least suspect it is a clear falsity or a very unlikely truth?17 Given the strength of this ‘suspect’, isn’t the resulting doxastic state very similar to, or even identical with, that of a person who entertains a pair of contradictory beliefs? Moreover, isn’t the project of self-deception, as intentionalists tend to describe it, self-defeating? That is, how can the deliberate intention of believing what one doesn’t believe, or it’s immensely hard to believe, succeed? If one knowingly decides to believe something, it seems that one cannot, for that very reason, accomplish one’s decision.

More precisely, the problem is that a belief is by its very nature truth-oriented and representative of the world. The state of affair the

\[16\] D. Davidson, Deception and Division, cit., p. 87.

Self-Deception

belief content represents determines the doxastic response of the competent epistemic agent. Evidence the subject faces contains an element of necessity which is constitutive of believing. The subject is not, as it were, ‘free’ in believing what the evidence backs up or points to. If I see that outside it rains, I immediately form the belief that it rains. Maybe my eyes do not work properly or there might be other, more or less far-fetched reasons why I should revise that belief. But it is indisputable that I perform my judgment and I acquire the belief immediately, no soon as I am presented with the extremely good evidence I have got when I look out and see that it rains. So, to have factual beliefs at all is to adhere to this element of necessity: unless one has not recognized the relevant evidence ‘as’ relevant – which seems to be extremely hard to happen in the examples presented – or already has some valid countervailing reasons for judging the evidence misleading, it seems that one can but immediately form the belief without further ado. Believing, thus, is not an act of free will18 and to be a believer at all is to be but incapable to escape the strength of the evidence once appreciated as such.

So, let us come back to what is inherently puzzling in having an intention such as that of bringing it about that one believes a proposition contrary of what the evidence one possesses justifies. Conceiving of such an intention is deeply irrational; indeed, one of the most irrational projects one could ever plan. Not only would it mean to shirk the concept of belief a rational agent possesses,19 since the belief thus formed would be responsive to the mere intention to form it, instead of to the evidence for it; it also means to misapply the concept of intention or to have, to say the least, a bizarre concept of it, since in order to have an intention I have to be prepared to take whatever steps is necessary to fulfil it; but in the case of ‘intending to believe’ the constitutive intention would intelligibly prescribe nothing else than intending to form a belief on a certain subject matter - that is to say, to attend the relevant evidence, which is precisely what from which the self-deceiver somehow must have deviated. Either one cannot conceive of any such intention - and self-deception must then be explained, or explained away, along non-intentionalist lines - or this intention must be stipulated to be of an unconscious kind, although active in a way

18 An exhaustive set of discussions about the present question are gathered in «The Monist», 85, 3, 2002. The volume is titled Controlling Beliefs.

19 That in order to have beliefs a creature must possess the concept of belief is widely discussed by Donald Davidson throughout his work. For wide argumentations in defence of the thesis, see A. Millar, Understanding People. Normativity and Rationalising Explanation, cit., chapter 1, 4 and 5.
that critics who advanced a similar hypothesis have traditionally not an easy time to clarify.\textsuperscript{20}

Notice that the main problem for intentionalism is ‘just’ the intentional character of self-deception, more than the possibility of being in an intractable, self-contradictory state of mind mentioned above.\textsuperscript{21} The latter possibility can be more easily dismissed by supposing, for instance, that the agent simply changed his mind after self-deception had been completed; or that the process did not get started by a belief, contrary to the ending one, which is maintained throughout the process and after it has been over. Equally, other problems it faces and that have been lucidly stressed in the literature, can be, as we will see, rather easily defused. But as far as an intention of bringing it about that I believe a proposition in the face of its supporting evidence is taken as central to self-deception, it still remains unclear how such an intention could ever succeed. It isn’t certainly enough to have, either consciously or unconsciously, such a project in order to accomplish it because, given the remarks made above about the nature of believing and intending, the remaining explanatory alternative would be that it be the case that people should count, at some point of their self-deceptive process, on some peculiar and, so to say, ‘lucky’ psychological mechanisms, deputed to guarantee the transition from their intention that p to their belief that p, for the project to be completed. The intention alone cannot make for its fulfilment.\textsuperscript{22} The likely verdict will well be that nothing like self-deception can ever possibly exist.

\textsuperscript{20} Sartre has famously criticised Freudian repression by pointing out that if the censor is in charge of pushing the painful belief out of consciousness, we have to suppose that the censor knows that he is bringing about the deception. Then, another censorial item must be posited, which would be in charge of removing from the censor the awareness that it is working to deceive the conscience. See J.P. Sartre \textit{L'être et le néant}, Gallimard, Paris 1943, trans. \textit{Being and Nothingness}, Hazel Barnes, Pocket Books, New York, chapter 2. The threat of the regress undermines the theory of unconscious intention. Others critical remarks on the idea are in J. Heil, \textit{Doxastic Incontinence}, cit.; A. Bird, \textit{Rationality and the Structure of Self-Deception}, «European Review of Philosophy», 1, 1994, pp. 19-38; and A. Mele, \textit{Self-Deception Unmasked}, cit., chapters 2 and 3.

\textsuperscript{21} That the intentional character of self-deception would be the major threat to a non paradoxical account of it is maintained also by H. Fingarette, \textit{Self-Deception}, cit., ch.2.

\textsuperscript{22} One counterexample to the present claim would be that of a boy who manages to deceive his future self by recording on his diary a wrong date for a meeting. He counts on the circumstance that he will certainly forget that he entered the wrong date. That way, he is sure that he will acquire a false belief when he will read that page of his diary. But we might resist to count the case as paradigmatic of self-deception, for the most interesting case of self-deception is one achieved via epistemic reflection over one’s present predicament and one that do not rest on devices like weak memory or time to succeed. The example is presented by B. McLaughlin, \textit{Exploring the Possibility of Self-Deception}, in \textit{Perspectives on Self-Deception}, ed. by B. McLaughlin – A. O. Rorty, cit., pp. 29-62.
Anti-intentionalists have a straightforward answer to the latter problem. It has long been rightly considered a virtue of their accounts to have shown that the intentionalist analysis is flawed by its imposing a paradoxical intention on a process that can be instead non-paradoxically accounted for without appealing to that very intention. On the anti-intentionalist view, the self-deceptive process would not consist of an impossible transition from the intention to believe that p to the belief that p; rather the process is initiated by a desire that p which leads, in ways to be specified, to the self-deceptive belief that p. Although not triggered by any deliberate intention of deceiving oneself,23 the human cognitive machinery is shown to contain purpose-serving belief-forming regularities that, pace the intentionalists, aren't (and needn't) be activated by any intentionally self-deceiving process at all. Roughly, having a desire that p, even though p is not sufficiently warranted, may, in many cases, do the job and bias our reasoning so as to lead us to form irrational beliefs. Empirical studies have been largely invoked to support the claim and to do justice to the implausibility of an analysis of self-deception carried out only by means of conceptual investigation.

I will explore in more detail in section 3 some representative anti-intentionalist responses. It will serve my argumentative purpose to show that, although the anti-intentionalists correctly fix the explanatory limits of the stronger intentionalist positions, to all appearances, they in fact have to bank (and indeed several do so, what they claim notwithstanding) on a minimal intentionalist element in order to work. Otherwise, the anti-intentionalist account stands on pain of:

1) not differentiating sufficiently between human and animal psychology, so as to end up including certain animal appetite-behaviour patterns among the range of self-deceptive phenomena, whilst self-deception can be appropriately ascribed only to creatures who are self-understanders and self-explainers;

2) giving an incorrect account of the alleged ‘sister’ phenomenon of ‘wishful thinking’, or misinterpreting it so as to end up assimilating it to self-deception;

3) running the risk of being so comprehensive to force us to incorporate, among the cases of alleged self-deception, also cases of motivational biased cognition not mediated by any reasoning at all, which

23 At least one example of how one could intentionally acquire a belief on a proposition he now disbelieves is exemplified by the so called Pascal’s wager, where an agnostic willingly decides to act as if God existed in order to acquire over time the belief that God exists. I will not try to analyse what a similar case, if successful, would amount to. The case is distant from the ones I have in mind for reasons akin to those expressed in the previous footnote.
we might resist identifying as self-deception (e.g., partiality, fanaticism, immediate emotional responses, «precipitate cases» of motivated believing, etc.);

4) rendering self-deception a passive phenomenon, unanswerable in any relevant sense.

At the same time, I take it that the anti-intentionalist account contains the clues for acceptably completing the picture: anti-intentionalist accounts tend in fact to stress, though in a different fashion from the one I favour and in view of different results, a characteristic failure of self-knowledge that becomes the proper object of regret the ex self-deceiver himself indicates as crucial in his having been just the ‘author’ of his self-deception - an author, we may concede, with some extenuating circumstances. The same failure is what an interpreter correctly appeals to when faced up with the task of making sense of the self-deceiver’s predicament. He may even realize that self-deception was highly predictable, or almost unavoidable, in the specific case, but he can’t help judging not only that it was among the self-deceiver’s epistemic possibilities to conduct his epistemic reasoning differently, but also that it was ‘in principle’ among his capacities to detect the specific failure of self-knowledge I will describe. It is about these possibilities that blame applies and it is on the fully legitimate attribution of these capacities that blame is grounded.

2. Intentionalist accounts of self-deception.

Intentionalism in its traditional version is generally defended by arguments to the best explanation, which in turn rest on a conceptual analysis of the term ‘self-deception’ primarily conducted by looking at its lexical constituents. What the term immediately suggests is the idea that self-deception is a kind of deception, perhaps similar in relevant respects to other-deception, from which it differs by what the prefix ‘self’ would imply. We are invited to follow the line of thought that any deception worthy of the name is an ‘action intentionally performed’. The prefix ‘self’ (not unambiguously) would serve both redundantly to specify that the action in question is self-conducted and that the person against


25 Mele distinguishes between lexicalist strategies, example-based strategies and mixed strategies. Normally, however one presents one’s examples in order to show how one’s theory applies to them: that is, one selects the example on the basis of what one already thinks self-deception is. Compare A. Mele, Self-Deception Unmasked, cit., pp. 5-6.
whom deception is performed is just the person who has conceived of it. In contrast with others-deception, where an agent A is the author of the deception of which an agent B is the victim, in self-deception the author and the victim will be one and the same person. Self-deception is then presented as an action, engaged for a reason which provides support to a special kind of intention: intending to believe a proposition which may alleviate anxiety or boost the subject’s self-image, that is, believing a falsity in sight of non-truth-oriented goals. Remind the cases sketched at the beginning: the rational belief in each (that my wife is unfaithful, that I have a cancer, that my son is guilty of a crime, etc.) threatens the fulfilment of these goals. Since the self-deceiver is, by assumption, capable of detecting the irrationality of the false belief, the view that the irrational belief is formed intentionally presents the formation of this belief as a consequence of practical reasoning. Appealing to practical reason is the best explanation in order to preserve the assumed integrity of the believer’s theoretical reason: self-deception is an outcome of a project that is undertaken by the agent in order to fulfil a desire.

The desire in question, however, may at best be viewed as a cause for the self-deceptive belief. It entirely fails to be a reason for it. Appealing to practical reason doesn’t explain yet how that can concretely happen. It needs then to be explained how that can happen. Which non-rational forces, if any, must be posited to act against one’s best epistemic judgment? Which hypothesis on the architecture of the mind must be assumed to account for irrational phenomena?

Davidson has given a seminal account of self-deception along similar lines – an account largely held as the traditional one and which is the standard locus of quotation from intentionalism for its especial directness. In the context of marking the difference between wishful thinking and self-deception, on which I will come back later, he claims that while wishful thinking presents a sequence in which a desire that p produces the belief that p, self-deception requires «the agent to do something with the aim of changing his view». In addition, not only must the self-deceived subject bring about his own deception, he also must do so intentionally. Davidson writes: «[…] it is not self-deception simply to do something intentionally with the consequence that one is deceived, for then a person would be self-deceived if he read and believed a false report.

---

26 See J. Heil, *Doxastic Incontinence*, cit., pp. 65-67, for a distinction between «one’s best epistemic judgement» and a judgments «all things considered», where not merely epistemic considerations can be weighed.

27 D. Davidson, *Deception and Division*, cit.

28 Ivi, p. 87, my emphasis.
in a newspaper. The self-deceiver must intend the ‘deception’. But how is this supposed to work? Here is Davidson’s elaboration:

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. […] 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.

There is clear textual evidence that Davidson’s view is that the two contradictory beliefs are both held consciously. Take the example he gives of Carlos, a guy who has good reason to believe he will not pass the test for a driving licence (he has failed twice, his instructor, on whom he trusts, has said discouraging things, and so on). He is aware that his total evidence points to failure. Other things being equal, he thinks it is better to avoid pain; the believing that he will fail the test his painful; therefore, it is better to avoid believing he will fail the test. «Since it is a condition of his problem that he take the test, this means it would be better to believe he will pass […] His practical reasoning is straightforward». But «core cases of self-deception demand that Carlos remain aware that his evidence favours that he will fail»; therefore, «Carlos believes inconsistent propositions if he believes that he will pass the test and that he will not pass the test».

A set of considerations is in order here fully to clarify what the view, as it stands, is essentially up to. Firstly, the account thus posed faces the problem of explaining the occurrence a paradoxical, indeed impossible, state of mind: a couple of contradictory beliefs are held contemporary by the same agent within the same awareness area. This is the ‘static paradox’ of self-deception. The trouble here is two-folded: how to

29 Ibidem
31 Compare Ibidem, p. 81, where he discusses the requirement of total evidence for inductive reasoning
32 Ibidem, p. 89.
33 Ibidem, 90.
34 Ibidem, 91.
explain the belief in the conjunction of the two contradictory propositions finally believed and how they can sustain each other, as Davidson claims at the end of the passage quoted above. If I could consciously believe at the same time both that \( p \) and that not \( p \), I certainly couldn’t fail to believe also their conjunction; that is, I would believe that \( (p \land \neg p) \), a contradiction.\(^{36}\) To avoid this problem, one might suppose that the two beliefs may well be inferentially insulated, and this is the solution Davidson goes for, by proposing the explanatory hypothesis of a certain kind of ‘partitioning’ of the mind.\(^{37}\) But then it is unclear how the two beliefs are supposed to interact on each other, if separated. A dilemma looms large: if they are not inferentially insulated they fail causally to interact on each other for rational reasons; if instead they are isolated they could not reciprocally interact for mere causal reasons. Furthermore, in either way we have at best a description of the state of mind in question, not an explanation of it. Also, in either way self-deception is explained away: not only do we lose touch with the agential manoeuvre it is claimed to consist of, but we also fail to make sense of how it could be achieved at all, given the dilemmatic situation described.

The second problem of Davidson’s intentionalism is in fact a dynamic one, the so-called ‘dynamic paradox’ of self-deception.\(^{38}\) We have just seen that it can’t intelligibly be the belief that not \( p \) to trigger the process which leads to the self-deceptive belief that \( p \). An intentionalist may then weaken the ‘two-beliefs requirement’ and allow for less strong theses. For instance, we may suppose that in self-deception a believer gets a change of mind: he believed that not-\( p \), but then after having negatively interpreted the evidence at his disposal, gathered new favourable evidence for the desired proposition, and so on, he is now convinced that \( p \). Still, intentionalism faces a problem: he intentionally engaged the process for the ‘reason’ of leading himself to believe that \( p \) and so for the ‘reason’ of changing his mind. He must intend his deception, that is, he knowingly takes up the project. This time around, he ends up in a consistent state of mind, but he must know that the final doxastic state has been produced

\[^{36}\] Notice that two propositions, \( p \) and \( q \), may be contradictory, while the beliefs that \( p \) and the belief that \( q \) may be inferentially insulated, as it happens in a suitably complicated theory which contains two inconsistent beliefs both believed by a theorist without him having yet spotted their inconsistency.


\[^{38}\] A. Mele, Real Self-Deception, cit.; and Id, Self-Deception Unmasked, cit., ch.3.
by himself. Any such intention, again, renders the process self-defeating. In other words, the desire to believe that p cannot intelligibly figure among the ‘reasons’ for the process to be engaged. It may at best figure among the ‘causes’ for it.

We are back to the cluster of problems we began with when cataloguing the alleged cases of self-deception: we have a description of irrationality, of the puzzling breakdown of reason (either theoretical or practical) it exhibits, still in need of an informative explanation as to «how it is done». 39

3. The ‘deflationary’ model

Anti-intentionalists generally argue for their position by starting from the intentionalist account’s intrinsic difficulties seen above. Their position is held to be ‘deflationary’ if compared with the intentionalist paradoxical descriptions of self-deception: self-deception exists, they claim, it can be accounted for in non paradoxical ways, and it has a straightforward explanation to the extent that we are willing to cart off the ‘intention to believe’ which intentionalists posit as essential to it.

Here are the four conditions Al Mele, the most influential among deflationists, indicates as ‘sufficient’, although not jointly necessary, to self-deception: 40

a) the belief that p which S acquires is false;

b) S treats data relevant, or at least seemingly relevant, to the truth value of p in a motivationally biased way;

c) the biased treatment is a non-deviant cause of S’s acquiring the belief that p;

d) the body of data possessed at the time provides greater warrant for not p than for p. 41

In particular, it is d) which is not necessary to entering into self-deception. The importance of d) not being necessary will be stressed below. Notice that no intention is cited in (a)-(d). Notice also that no two-beliefs requirement is mentioned as well.


41 See Id., 2001, Self-Deception Unmasked, cit., ch.3.
The basic idea contained in (a)-(d) is that a desire that p be the case is sufficient to trigger a number of biases affecting our cognitive process to the result that we can arrive at believing a falsity. Examples of the motivational biases Mele refers to in spelling out the condition b) are: selective attention to evidence that self-deceivers actually possess; selective means of gathering evidence; negative misinterpretation (failing to count as evidence against p data that we would easily recognize as such, were we not motivationally biased); positive misinterpretation (counting as evidence for p data that we would easily recognize as evidence against p, were we not motivationally biased). Some other biases Mele introduces are presumably operative also in ‘cold’ instances of irrational believing, that is to say, when no motivation is at work, as vividness of information, availability heuristic, confirmation biases.42

As to (c), the desire in question is a non-deviant cause in that it is ‘thematic’: it is just the desire that p be true that leads to the belief that p via the processes described. The self-deceiver is content-sensitive to the desire that p and his content-sensitivity is relevantly explanatory of his self-deception. More complex scenarios can be of course figured out, where less direct desires are thematically linked to the desire that p be true or to the belief that p, as for instance in the negative, or ‘twisted’ cases of self-deception, where S comes to believe that p because he feared that p be true (e.g., a husband unjustifiably believes that his wife is unfaithful because he fears that she is unfaithful).43

Self-deception is no more than a process involving the above ingredients, which are normal players of our general psychological life. Nothing special, exotic or paradoxical, no ad hoc psychological item or process needs to be evoked in order to make sense of self-deception.

I will direct in a moment my attention to what ‘desiring that p be the case’ amounts to, whether this desire needs be aware and how it conceptually differs, if any, from ‘desiring to believe p’. For the time being, let me present in more detail what the connection is, on Mele’s view, between the desire that p be the case and the exercise of motivational biases. Mele embraces44 the model of everyday hypotheses testing...
developed by Trope and Liberman\(^45\) as the empirical demonstration of how motivation can bias cognition. The story goes as follows: people have different acceptance/rejection thresholds for hypotheses depending upon the expected subjective cost to the individual of false acceptance or false rejection relative to the resources required for acquiring and processing information. The higher the expected subjective cost of false acceptance, the higher the threshold for acceptance, and similarly for rejection. Hypotheses that have a high acceptance threshold will be more rigorously tested and evaluated, while those which have a low acceptance threshold will be more quickly embraced. Now, in self-deception the expected subjective cost associated with the acquired false belief is low. For example, the husband who falsely believes that his wife is faithful has different acceptance thresholds for the hypotheses of falsely believing that his wife is ‘not’ having an affair and for the hypothesis of falsely believing that she ‘is having’ an affair: he is much happier at falsely believing the first hypothesis than at falsely believing the second, because he desires that she is not having an affair. Clearly, then, Mele has it in mind that S’s desire that p be true results in a motivationally biased treatment of data by lowering the acceptance threshold and raising the rejection threshold of the hypothesis that p, thus opening the door to the biased treatment.

It is worth noticing that the operative notion of this account is that of ‘hypothesis’, not of ‘belief’, as in Davidson’s account discussed above. This largely helps clear the ground from the difficulties surrounding the traditional intentionalist strategy already unpacked. Taking a hypothesis that p in order to test p’s truth value is akin to the attitude of entertaining the thought that p, or suspecting that p. Self-deception needn’t begin from a stable belief that p. Accordingly, if self-deceivers do not believe that p, but merely suspect that p might be true, they do not engage in a process which would hopelessly make violence to their doxastic steady convictions. At most, they may begin with believing that p, but then take up undermining the unbearable belief and so come to change their mind by means of the same mechanisms deputed to lowering the acceptance threshold of the proposition that p. That is, the belief that p is momentarily revoked and now the hypothesis that p is in place and available to be tested. It is intelligible that one person under high motivational pressure can ask herself whether she should go on believing that p. If the self-deceptive manoeuvre will be successful, she will have changed her mind. No static paradoxes need occur.

But let me focus on the anti-intentionalist core claim, namely the idea that self-deceivers do not intentionally bring it about that they believe that \( p \), further to the desire that \( p \). We must be careful, I think, to what desiring to believe that \( p \) can intelligibly mean.

If a hypothesis is at work, and if the agent has any interest at all in testing its truth conditions, this means that the agent cannot fail intentionally to test it. We can say, in this sense, that he ‘intends to form a belief relative to \( p \)’. In particular, when motivationally biased by the desire that \( p \) be true, one intends to establish that there is sufficiently cogent evidence that \( p \) is true. That way, one can form the belief that \( p \). This does not resemble any puzzling process like intending to believe \( p \), as it were, ‘at will’. Self-deceivers do not make up their mind so as to ‘decide to believe’ that \( p \). In the latter case, we will be facing the intrinsic difficulties of intending to believe seen in section 2. In the former case, instead, no such happening needs be encompassed. Deceptive desires operate towards the ‘fact that \( p \)’, not towards the ‘fact of believing that \( p \)’ (where the former is a fact about the external world, whilst the latter is a psychological fact about the agent). Self-deceivers normally do not want to acquire a state of mind for its own sake, but rather the certainty, or at least a degree of assurance, that \( p \) is true. Given the nature of a hypothesis, then, it seems intelligible that one intentionally proceeds to test it. One may fail to recognize that one favours the hypothesis that \( p \) over not \( p \) (I will come back to this in section 5), but not that one is working to test the hypothesis. Anyone who has acquaintance with his/her own or others’ self-deception may recall how the self-deceiver overtly evaluates his or her hypothesis and how fervently wants to find evidence for \( p \). I myself find it difficult to think of a clear example of motivationally biased hypothesis testing which is not intentional in the sense described. So, although Mele correctly presents his account as anti-intentionalist (no intention to deceive oneself) if compared with the traditionalist intentionalist position, he does not do justice, and rightly so, to a minimal intentionalism of the kind described.

That this be correct, it is also suggested by the fact that others deflationists, Ariela Lazar in particular, recently object to Mele to fail to be a thorough deflationist. Let us then see how other versions of anti-intentionalism work and how they fail to give a sufficiently specific


account of self-deception. Subsequently, I will say how Mele account suffers from a flaw of its own, although it is open to a critic to interpret his account as minimally intentionalist. Since I think that a minimal intentionalism is what we need to capture the phenomenon of self-deception, I will need to show that there are reasons, independent of his minimal intentionalism, which debar Mele’s (a)-(d) conditions from being - again - sufficiently specific of self-deception. The d) be not necessary will be crucial to my argument. The entire discussion will eventually set the stage directly for the final analysis of the view I favour (section 4, 5 and 6), and its corollaries on human rationality, failure of self-knowledge and blame.

Lazar claims that Mele fails to avoid the problem of «crazy choices» affecting intentionalism in the first place and Mele’s (only nominal) anti-intentionalism, too. Let’s recall, as a paradigmatic example of what Lazar has in mind, the case - presented at the beginning of this paper - of the skilled oncologist, call her Ann, who acquires the false belief that she is healthy despite the evidence suggesting that she is probably not. If Ann’s self-deception is to be explained as driven by the strong desire of leading a long and healthy life, it is puzzling that some critics end up – as Mele would do – appealing to the desire to believe that such life is forthcoming. The puzzlement is generated by the fact that these two goals will clearly be in conflict: in Ann’s case the fulfilment of the goal to believe that she will lead a long life is obtained at the expense of satisfying the original desire, namely, to lead such a life. In other words, the question of whether the false belief in self-deception is formed by the agent for the ‘reason’ of wanting to form that belief is crucial and while Mele rejects the view that the self-deceptive belief is formed because of an intention to form it, he nonetheless suggests that it is the ‘desire to believe’ that accounts for the formation of the irrational belief. The point is conceptual: if the ‘desire to believe that p’ is operative in self-deception as a ‘reason’, instead of being the ‘desire that p be the case’ merely the operative ‘cause’ of the whole process, we have self-deception reformulated as a paradoxical practical sequence: self-deceivers’ choice does not make sense by their own lights, since they can’t fail to be aware that the desire to believe that

48 She presents three objections: ‘negative (or twisted) cases’, ‘how it is done’, and ‘crazy choices’. The second one has already been made clear. The first one is hinted at in footnote 43.


50 Compare A. Mele, *Self-Deception Unmasked*, cit., sec. 4, par. 8 and 10.
p cannot be an adequate instrumental reason in the service of fulfilling of the original desire that p. It is Lazar’s view that the desire that p suffices to trigger self-deception: we needn’t suppose that any such practical, mediated reasoning is required to self-deception. Rather, it is just the adoption of such a picture that makes the case against alleged «crazy choices» in self-deception.

I think that Lazar's criticism to any such explanation of self-deception is fundamentally correct, but I doubt that Mele's minimal intentionalism entirely matches the feature of that sort of explanation. This is because on the hypothesis testing model ‘desiring to believe’ is prima facie suggested to mean – as I have already pointed out – desiring to form a belief about the hypothesis that p. Unbeknownst to the self-deceiver, the desire that p be true causally biases his epistemic strategy, as Mele's condition b) clarifies. This does not entail, however, that the self-deceiver intends the self-deceptive belief to be instrumental to fulfil a desire. Ann has a desire, which may be even consciously held, to be healthy. She then evaluates if the hypothesis that she is healthy can be taken as sufficiently justified. She looks for evidence confirming that she is. She interprets the evidence at her disposal and seeks new one in order to establish that it is true that she is healthy. The epistemic endeavour she undertakes is not 'so crazy' after all; certainly it isn’t crazy in the sense described by Lazar. Granted, Ann intends to form a belief about p; she does this in a motivationally biased way; but at no point of the testing process described by Mele she believes that it is better for her to believe that p as a means to satisfy the desire that p be true. However, I concede that the question raised by Lazar may still be taken as open. To proceed to a textual investigation in Mele's pieces of work will lead me too far. I am in fact in this essay less interested in adjudicating the present dispute between Mele and Lazar than I am in explaining why full anti-intentionalism, as Lazar's and Johnston's, fails to capture self-deception as a specific phenomenon among the wider range of occurrences of motivated irrationality. When eventually the difficulties of anti-intentionalism will be made clear, Mele's failure to be a thorough anti-intentionalist will well be available to be cashed out in terms of an advantage of his position over other deflationist accounts.

Now, anti-intentionalists like Lazar and Johnston point to account for self-deception by deflating not only the role of the traditional intention to believe a falsity, but more generally the role of reflective reasoning altogether. They aim at dismissing what they take to be an over-intellectualist picture of the mind, mainly due to Davidson, affecting the explanation of self-deception as well as the explanation of belief-formation under the influence of emotions on the whole. On their views, distortion of cognition by interest or desire is achieved by means of mechanisms
that do not encompass rational exercise and conceptual mediation on the part of the self-deceived believer. By supposing that self-deceivers do not entertain any second-order belief about what they ought to believe or to what end, Lazar and Johnston aspire to avoid ‘both’ the problems of doxastic incontinence (believing in the face of evidence), which compromises the believers’ theoretical reason, ‘and’ the problem of crazy choices (believing a falsity as a means of fulfilling a desire that that very belief in fact frustrates), which compromises the believer’s practical reason, without substantial explanatory gain in either model.

Here is Johnston:

The sub-intentional mental process involved in wishful thinking and self-deception is an instance of a non-accidental regularity: anxious desire that p, or more generally anxiety concerning p, generates the belief that p [...]. Hence I speak of a mental tropism, a characteristic pattern of causation between types of mental states, a pattern whose existence within the mind is no more surprising, given what it does for us, than a plant’s turning toward the sun.51

Mental tropisms are, according to Johnston, non-accidental, purpose-serving mental regularities between types of mental states, e.g., desires and beliefs, that are not mediated by intentions to act. They are somehow akin to unreflected habits or reflexes of the mind. In Johnston’s words, they are «blind but purpose-serving connections»52 between desires and beliefs. Self-deception is then a process where the desire that p and an anxiety that not p generates the belief that p, even though the agent has evidence to the contrary. Such mechanisms are operative in us because of the vital role they perform: reducing anxiety, which in turn gives us survival advantages. We certainly have a certain control over our anxiety, so we have to suppose that the self-deception tropism is activated when we fail to exert such control.

Ariela Lazar presents an account of self-deception53 in the same spirit of Johnston’s. The self-deceptive beliefs are, on Lazar’s view, a direct result or expression of the self-deceiver’s wishes, fears, and hopes. She calls those beliefs «a kind of fantasy».54 The effects of these emotions and desires are not mediated by practical reasoning. No negotiations with the agent’s epistemic standard need to be in place: «Emotions do

52 Ivi, p. 88.
54 Ivi, pp. 284-287.
not affect one’s view of the world through deliberation; they do so immediately and in a way which, to a high degree, is not subject to our control.\textsuperscript{55} […] the formation of emotion-induced (or desire-induced) beliefs is highly uniform, quick and effortless.\textsuperscript{56}

The trouble with both accounts is that if there is a notion of self-deception worth having, it must give us more than just being badly mistaken due to strong emotional affects, anxiety and the like, even if this is what Lazar thinks self-deception amounts to: «We label self-deception those beliefs which are at odds with the evidence and which are highly influenced by the presence of emotions and evaluative attitudes»,\textsuperscript{57} and even if Johnston conclude that: «To be self-deceived is sometimes just to be misled without being intentionally misled or lied to. The self-deceiver is self-misleader».\textsuperscript{58}

I take that Mele does not seem to deny the role of reasoning altogether, given that he adheres to the model of hypothesis testing; still, the trouble with his account is that he denies that reasoning is a necessary condition to self-deception. Indeed, it is intelligible that one test a hypothesis blindly, hastily and that the testing operation thus be, or be next to, a one-step operation which tends to lose the character of a reflective, though interested, hypothesis testing. If the hypothesis that $p$ so rapidly is installed as the belief that $p$, which is entirely possible, we face a case of motivated irrationality, but not as much clearly as of one of self-deception. The clue to understand that this be so is clarified by the circumstance that condition d) in Mele’s set of sufficient conditions (a)-(d) for self-deception is claimed not to be necessary to self-deception. If condition d) is not necessary to self-deception, and if self-deception doesn’t essentially require a reasoned hypothesis testing, the unwelcome result might well be that an animal could easily be counted as self-deceived, on a deflationary account, like Mele’s. By the same token, human prejudices, partialities, wishful thinking and precipitate cases of believing will largely overlap with self-deception, while there are reasons for a similar overlap not to occur and for keeping the phenomena distinct.

Let me then introduce the case through the following scenario, drawn from Scott-Kakures,\textsuperscript{59} which contains the first of the four critical points

\textsuperscript{55} Ivi, p. 282.
\textsuperscript{56} Ivi, p. 283.
\textsuperscript{57} Ivi, p. 287.
\textsuperscript{58} M. Johnston, \textit{Self-Deception and the Nature of Mind}, cit., p. 65, emphasis by the author.
I advance against strong anti-intentionalism. They all have been previously mentioned (section 2, p. 155-56): the need of differentiating between human and animal biased cognition on the one hand, and specifically human self-deception on the other; the need of differentiating between wishful thinking and self-deception; the need of differentiating between precipitated cases of believing and self-deception; the need of doing justice to the active role of the self-deceiver in accomplishing self-deception, even once we reject the traditionalist intentionalist picture.

1) Why an animal cannot be a self-deceiver.

Scott-Kakures presents the following case in order to test Mele’s deflationism about self-deception. Consider Bonnie, the cat.

Like most felines, Bonnie can make fine aural discriminations. She can, for example, distinguish the sound of the removal of her own medication from the cupboard from the sounds of the removal of other objects – she promptly disappears only when her own medication is removed. Bonnie is also exceedingly fond of her food. She is apt to scamper into the kitchen when she hears a can of her food being opened. She rarely so scampers into the kitchen when some non-cat-food item is being opened.60

Yet, on occasions upon which Bonnie is very hungry she mistakes non-cat-food sounds for cat-food sounds, whilst in the absence of hunger she doesn’t make such mistake. We may assume that Bonnie enjoys some form of belief-like and desire-like states, although it is reasonable to suppose – and indeed this is the point on which Scott-Kakures makes his argument trades - that these very states are different in kind from humans’. We may describe the situation Bonnie is in so: «When Bonnie is under the influence of a strong desire for food, she is apt to come to believe falsely that the cat food is currently being opened in the kitchen»61. Should we say that Bonnie is self-deceived? On Scott-Kakures’ opinion, a deflationist like Mele has scant basis to deny such a characterization. Recall Mele’s (a)-(d) conditions: firstly, Bonnie’s belief is false (a); secondly, the desire is a non-deviant cause of the belief reached (c); thirdly, the data she possesses are treated in a motivationally biased way (b), and even though we might grant that she does not treat data in the agential manner a human being treats factual judgments and normative epistemic principles, she is nonetheless an information processor, whose operation is affected by a desire in the present case. Furthermore, it is in the very

60 Id., p. 578.
61 Ivi, p. 579. For another explanatory hypothesis on the case, see Charles Young’s suggestion, reported in ibidem, footnote 7.
nature of a bias to get executed beyond any agential control over its activation. Sometimes we cannot avoid being biased, as it is shown in the literature about cognition and cognitive unconscious, although we can get trained not to persevere on keeping the bias up once trained to detect it as such. That this can be so for humans renders even more intelligible the case for animals, which certainly do not enjoy our conceptual sophistication over their mind.

That condition (d) is not relevant to exclude Bonnie’s being self-deceived is granted by the fact that (d) is not necessary to self-deception: the fact that the self-deceiver recognizes the strength of the unfavourable evidence at his disposal and that, on that basis, he goes on testing the favoured hypothesis through methods which do not meet his best epistemic judgement may be sufficient to self-deception, but not indispensable, on Mele’s view. Notice that in self-deception embarking on the hypothesis testing process makes sense only if we assume that the self-deceiver recognizes that some feared ‘propositions’ are more likely true than false. Were not the proposition feared and were not the evidence recognized as appropriate to conclude as to its truth value, the self-deceptive process would presumably not get started. Bonnie cannot certainly undertake any such process: she doesn’t possess the conceptual complexity necessary for second-order beliefs and reflection upon her doxastic condition which (d) points to. This is in fact what sanctions the difference in kind between humans and animal doxastic and psychological life.62 In any case, I do not want to exclude that frequently we reach doxastic conclusions without employing much sophisticated strategies. The all point I need to make is to contend through the fours argument I have been giving for these phenomena deserving the name of self-deception.

In sum, however counterintuitive it may appear, if (d) is not assumed as necessary, nothing prevents, on the deflationary model, to include Bonnie among the paradigmatic examples of self-deception. By the same token, human self-deception itself risks to fade into wishful thinking, as I will say in the following set of considerations.

2) Why self-deception is not akin to wishful thinking.

It isn’t immediately clear whether or not wishful thinking should be taken as relevantly similar to wishful believing, but we can certainly consider as empirically real cases where a person’s indulging in wishful thoughts,

---

62 Clearly, we are to suppose that the first-order beliefs of animals are different in kind from human ones in the first place, if anything because animals lack language and it is extremely hard to say that they make factual judgements relevantly similar to ours.
her entertaining desirable fantasies which express her own wishes, fears, emotions and the like can influence her behaviour as much as beliefs do. This is the claim Lazar puts forward\textsuperscript{63} when explaining how self-deceptive beliefs are a kind of fantasy and how self-deception should be best viewed as a kind of wishful thinking gone wrong. The phenomenon is irrational: the believer should conduct herself better than she does, by avoiding assuming her fantasies as evidence of her enjoying certain desirable features, or people being as she can’t help seeing them to be, or certain state of affairs being undisputable facts. It is clearly also motivated. While wishful thinking in itself cannot necessarily be irrational (there is nothing immediately irrational in fantasizing as such), wishful believing is more clearly on the irrational side. Crucially, however, human full-blown self-deception operates in the face of contrary evidence at the believer’s disposal. On the contrary, wishful believing can entirely be accomplished without any unfavourable evidence in place and so without any epistemic work on the part of the believer. Suppose I believe in God since I was a child. I have always thought of God. The thought that God exists is wishful for me and I have never displayed any interest in determining whether or not God exist. I obviously lack any conclusive evidence over God’s existence, but I do not care at all about the evidence. I can even feel no need to give an answer to the atheist’s challenge: I am completely deaf to his reasons, maybe I have never been sufficiently closely exposed to his argumentations. One might say that I believe in God ‘because’ the thought of his existence is overwhelmingly wishful, but the order of the explanation should be reversed by my lights: I believe in God through education ‘and’ the thought is wishful. Do I count as a self-deceiver? Maybe I am merely ‘self-deceived’, intending with the past participle that I have been ‘misled’ by education and exposure to a deeply religious social environment. I have acquired a belief about the proposition that God exists without testing it in any relevant sense, without ever reasoning on it and without ever striving to prove its reasonableness. The case will of course be different if, as a faithful, I began to doubt at some point that God exists and I engaged in a process of ‘interested’ justification of my belief in order to retain it.

A similar point could be made to explain cases of optimism, pessimism, prejudice of various sorts, non-painful cultivation of self-image,\textsuperscript{64} and so on.

\textsuperscript{63} A. Lazar, Deceiving Oneself or Self-Deceived? On the Formation of Belief ‘Under the Influence’, cit.

\textsuperscript{64} One beautiful example comes from French literature: the complex cultivations of beliefs about herself that the main character of Gustave Flaubert’s \textit{Madame Bovary} magnificently manages to attain. G. Flaubert, \textit{Madame Bovary}, Lowell Bair, 1967, trans. Bantam, New York 1959.
Wishful and fearful thinking is synonymous of delusion, fantasy, day-dreaming. If it goes wrong so as to turn into wishful believing and influence our behaviour, this may be explicated by appealing to partiality: I have never taken into consideration what I’d better think about the belief that p, or I have never had any occasion on which my attention has been drawn to reconsider that maybe it is unreasonable to cultivate a certain prejudice I have got and which has always directed my related evaluation. In a word, I have never been called upon taking a second-order stance on my belief that p. I may well be criticised for not being sufficiently critical or intellectually alerted, even for being self-deceived, but not for being a self-deceiver.

3) Precipitate cases of believing vs. self-deception.
Lazar presents the case of John, a man who lost his job. He feels depressed and at times believes that he will not have any chance to get a new job. However, after a dinner spent with his best friends, he feels reassured and he now believes that he indeed has good chances of getting a new job. Lazar gives the examples as one which would show how emotions and feelings, along with desires, are normal players in the belief-formation process and that self-deception doesn’t require more than this in order to succeed. No doubt the plausibility of similar belief occurrences should not be played down: we are all acquainted with such doxastic seesaw when heavily distressed. Emotions largely prevent our doxastic life from following our best epistemic standards. The point I wish to stress, again, is whether this suffices to self-deception. Isn’t it more suitably describable as a case of ‘precipitate believing’? Isn’t the case described one in which no significantly accurate epistemic work was done in order to reach the favoured conclusion? The man does not weigh ‘all’ the evidence at his disposal, maybe after the conversion with his friends he entirely forgot the negative evidence he attended before meeting them. Conversely, he had formed the pessimistic belief without thinking at all of the countervailing considerations coming from the thought that he has dear friends who trusts on him, despite the recent professional failure.

67 Lazar avoid saying what the conversation John has with his friends concerns. Maybe they directly provide him with new evidence about his chances to get a new job. It remains clear, however, that even though they do not discuss about John’s unemployment, John can still use their love and make a number of inference from it, in order to derive the conclusion that he is not hopeless as a man and, consequently, as a worker.
His epistemic standards are not operative in their fullest either before or later. He needn’t work on them in order to back up his desired conclusion. That emotions, pleasant thoughts, and pleasure in general shape our view of the world does not entail that we are self-deceiver, although we may result even badly mistaken when in the grip of them.

4) Self-deception as an active, agential process.

I have argued thus far that the peculiar epistemic activity the self-deceiver deploys marks the difference between self-deception and the cases of motivated irrational believing sketched above. In precipitate cases of believing we assist at a suspension of the normal epistemic standards of judgments – a suspension we frequently cannot avoid, for instance when under the influence of a strong emotion. The case of wishful/fearful thinking equally does not display any significant reasoning activity on the part of the believer and it is akin to cases of partiality, prejudice, etc. Such believers are irrational because they make their inference so hastily and inaccurately, they simply ‘jump to the conclusion’.

The role of reflection upon one’s doxastic predicament and a laborious epistemic negotiation of a kind to be specified about what we rationally ought to think of a certain proposition we fears or desires given the evidence is the mark of self-deception. No such lucid negotiation is necessary to make sense of the cases of motivated irrationality just seen. No relevant agential intervention on the part of the believer is required to shape the direction of their cognition. However irrational their doxastic behaviour and odd their conclusions may appear, they fail to have the degree of control over their epistemic strategy the self-deceiver shows to enjoy.

Before I proceed toward my view, I wish to precise that I am not contending for the diagnosis of self-deception of any of the cases thus far quoted in particular, nor do I wish to say that those I presented as cases more difficultly gatherable among the ones of self-deception must be not. Self-deception is an interpretative hypothesis over real cases of false believing and in many circumstances it can be difficult to judge how the process has been really achieved. Maybe, most of the alleged cases of self-deception in the literature (as those presented at the beginning) are best accounted along the lines of precipitate believing. In general, what I wish to vindicate is the conceptual and empirical possibility of much more complex, though non paradoxical, cases where the intention to test the favoured hypothesis shapes cognition through means of reasoning and full-blown sensitivity to reasons. In a word, where the believer ‘advocates’ for the desired conclusion as a lawyer would do, and in a way that at a remarkable degree satisfies the demands of his best epistemic principles.

Self-deception paradigmatically begins from a tendentious ‘practical question’:68 «What should I think about p?», e.g., «Does my wife really cheat me?», «Am I really ill?», «Do I really love that man?», «Did I really hasten my alcoholic husband’s death?». The installation of the hypothesis p is certainly the expression of the desire that p be true. This is, I think, a genuine bias the self-deceiver may be not aware of, while the self-deceiver may even overtly recognize that he wishes that p. That is to say: he may even conceptualise that he wishes that p, while failing to recognize that the wish has casually installed p in the guise of a hypothesis ready to be tested for confirmation (recall the confirmation bias and how motivation works on it, explained by Mele, see section 3). Our betrayed husband may say to himself, or to others, spontaneously (as an answer to a first critical evaluation of his own desire to test p) or when questioned: «It’s clear that I wish that my wife is faithful. That’s why I want to be sure that she is!». So, we needn’t suppose the desire to be unconscious, though operative, for self-deception getting started, whilst we may reasonably suppose that our mind generates the hypothesis itself without any control on our part over it and without any awareness that we are undergoing a similar bias.69

That being correct notwithstanding, we get little advantage in deflating self-deception up to the point at which it fades into the slumber of reason. Rationality, and the thirst for it, more than its sleep, generates self-deception.70 Real and common as they may be, other forms of motivated irrationality do not resemble the higher-level type of motivated irrationality that self-deception comprises. Self-deceiver’s conceptual sophistication is all in the service of the desired conclusion and in the ideal case of an epistemically and psychologically ‘very’ skilled self-deceiver who does no fail to realize that at every step of his investigation over p he needs to justify what he comes to think, successful self-deception can require time, great reflection and a considerable number of justificatory stories to pin down the self-deceptive doxastic conclusion. A chain of supplementary self-deceptive plots may be necessary to back up the conclusion. Most of them will not concern the status of p given

69 See A. Mele, Self-deception Unmasked, cit., pp. 30-33.
the evidence, but the value of the evidence; the reasons why not to seek
new evidence; the reasons why one is sometimes licensed not to apply
his best epistemic standards; the reasons why the situation the self-
deceiver is in can be grouped out among the exceptional cases; and so
on. In a word, self-deception is a vast and intense operation of «inventions
of reasons»,\(^71\) which often mobilizes our more sophisticated sensitivity
to what reason demands. Self-deceivers wants their beliefs to be justified
and true. As every reflective believer, they value having true beliefs.\(^72\) If
necessary, they may even proceed to undermine and sophisticate the sense
of what it is to have a belief. Sartre’s analysis of «bad faith»\(^73\) is focussed
precisely on such a radical, though not uncommon, I think, possibility.
Although the topic would deserve a long, separate discussion, I will try
to summarize what Sartre has in mind and I will restrict the discussion to
the introduction of the ideas that are immediately relevant to the present
purpose.

Owing to the unfortunate circumstances of our lives, we often fail
to be in the enviable position of the perfect knower. ‘Imperfections’ of
belief, in its etymological sense, is a normal predicament for human beings.
Arriving at forming beliefs on a host of subject matter is a project that
often fails, which falls short and is consciously left incomplete. «When
this happens», as Wood has put it in his discussion of Sartre’s bad faith,\(^74\)
«what we are left with as beliefs are things that are made to do the job of
belief, but which we recognize as insufficient to do this job». Wood
proposes then to interpret Sartre’s suggestion that in bad faith I
paradigmatically believe what I disbelieve in terms of a distinction
between two nuances of the concept of believing: I believe what is self-
evident and undeniable on the basis of the evidence presented to me,
but I recognize that I also believe (Wood calls it «to believe*»)\(^75\) on the
basis of inadequate, mixed, or ambiguous evidence. I accept to believe*
at times because, though I lack unequivocal evidence, I still need an
established way of dealing with the world and reacting to it (including
verbal behaviour, tacit speech, explanatory and practically orientating
thoughts). From the standpoint of the project of believing, believing*
results unsatisfactory, but we all are used to live with it.

Now, the skilled self-deceiver can favourably exploits this rift at the
heart of our concept of believing while carrying out his reasoning toward

\(^{71}\) D. H. Sanford, *Self-Deception as Rationalization*, cit., p. 158.
\(^{72}\) Compare M. Forrester, *Self-Deception and Valuing Truth*, cit., p. 44.
\(^{73}\) J.-P. Sartre, *L'être et le néant*, cit., ch. 2.
\(^{75}\) Ivi, pp. 217-8.
the self-deceptive doxastic outcome. There may be a point at which he overwhelmingly can’t refrain from recognizing the insufficiency of the evidence gathered or positively/negatively misinterpreted, the extremely weak character of his already performed justification not to seek new one so as to adhere to his best epistemic standards (like the requirements of total evidence for inductive reasoning). He needs a further justification as to why what he obtains after having tested his overall hypothesis testing strategy is ‘actually’ enough doxastically to conclude his reasoning and come to a halt: this justification lies in the disbelief about believing. He can end up saying: «No one can completely know everything all the time, I am aware in advance that I cannot do more than believe in many cases and for all I know this may well be one of them». I myself recall more than one time at which I have taken a stance of this sort relatively to highly thematic questions of my life by their very nature at risk of turning into self-deception, and I did not always manage to avoid reaching the false conclusion I later retreated. The reader is invited to try the experiment with the examples of self-deception previously given: the betrayed husband, the ill oncologist, the affectionate mother, the girl in love, etc. If they are not cases of precipitate believing, tropism-like jumping to hasty conclusions, ingenuity or ignorance cases, they all can be intelligibly enriched so as to discover in the self-deceivers’ train of thoughts similar epistemic considerations. I do not claim that they all ‘have to’ be so construed. I simply claim that they can.

5. What I do not know of myself qua self-deceiver

On the view of self-deception I have been presenting, the self-deceiver does not fail to recognize the he desires that p be true, nor does he fail to be aware that he is testing p for strengthen the evidence apt to justify p. He intentionally tests p, though he certainly does not intentionally test it in order to deceive himself. But then, what does he fail to realize? Where does his failure of self-knowledge, if any, lie? I grant him to be rational and epistemically competent, or he would not have been able to argue so finely for his desired belief, which is arduously reached and watched over. I also grant that, were it not be for the desire that p be true, the whole rationalising process would have not taken place. Something must have escaped his attention, for we are assuming that it was in principle and in practice possible for him to direct his cognition differently, toward the correct belief. I think that what escaped their epistemic control is the connection between the desire and the process that the former has triggered and sustained. They fail to judge the correct causal order between their attitudes, so missing the possibility of refraining from the obstinate search for justifying reasons for the
improbable proposition \( p \). Self-deceivers noticeably misconceive what animates their doxastic and cognitive activity. A similar view is worked out by Annette Barnes:

[...] the self-deceived will always misapprehend, will always not make a high enough estimate of the degree to which anxious desire that \( p \) contributes to the belief that \( p \). He or she will believe that the belief is justified, that the belief is based on evidential beliefs or on direct sensory perception to an extent sufficient for justification.\(^{76}\)

Notice the presence in this account of the various second-order beliefs associated with the epistemic status of the object-level belief, e.g., the belief that the belief that \( p \) is justified or that it is sufficiently grounded in evidence. In a word, the kind of beliefs their epistemic manoeuvre at every step requires in order to be carried out so as to result satisfactory to a rational creature. Scott-Kakures writes:

Like any reflective reasoner, [the self-deceiver] will regard her investigation as directed by herself, by her grasp upon what reason recommends; her search is a search for reasons that will permit her to bring satisfactory closure to her investigations. Yet, quite independently of her own evaluations, judgments, and activities her investigations are directionally driven by desire and interest.\(^{77}\)

Since the self-deceiver then exerts meta-cognitive control over the process responsible for the installation of the belief, minimal intentionalism is right: self-deception demands that there be an agent. At the same time, deflationists are right at claiming that self-deception is not intentional in the way described by the dynamic paradox, and that it is not a process at the outer limit of our understanding.

6. Failure of self-knowledge and blame in self-deception

We may wonder why a self-deceiver fails to judge the process he undertakes as highly suspicious. After all, not every believer in the grip of a desire that \( p \) be true ends up believing a falsity. Not only doesn’t the desire in itself suffice to trigger self-deception - this is what Bermúdez objects to Mele, by advancing a «problem of selectivity»\(^{78}\) for self-deception


Self-Deception that deflationism does not accommodate: many people refrain from similar search for reasons in the service of a desired conclusion; but it is also in many cases correctly judged as being the possible cause of self-deceptive temptations. I think that we cannot reasonably expect people to be invariantly able to reach such a refined, self-critical psychological judgment over themselves and their cognitive styles. Although their thirst for rationality is not suspended in self-deception, as it is in wishful thinking or other immediate forms of motivated irrational believing, and although they show to be clever reasoners, the bootstrapping of the self which would be necessary to avoid self-deception might still not be available to them. We have here an elucidation of the problem of selectivity: selective is one’s personality and the counterfactual circumstances that help shape one’s disposition to self-deception. If A desire that p, B desire that p, and A goes on deceiving himself, while B does not, their psychological situations were presumably different from the beginning. The desire that p be true does not suffice to trigger self-deception, unless others features, concerning the cognitive style of the person as a whole, are operative. Contrary to Mele, desire is necessary to self-deception, though not sufficient.

It is worth noticing that Forrester makes a list of epistemic virtues, as it happens, that can help overcome or avoid self-deception: firstly, a self-deceiver must believe that knowing the truth – in his situation and in general – will lead to his greatest overall satisfaction, and he must have some desire to know the truth. Secondly, he must be able to recognize self-deceptive tendencies – signalled primarily by his realization that […] he suspects that p might be true […]. Apart from seeing self-deception in himself, he surely knows that in general people tend to believe what they want to believe and thus could be alert for this trait in himself. Finally, he must manipulate his motivation so as to diminish his desire to avoid the truth.79

That one lacks the virtues in question or the fact that one may well not be sufficiently acquainted with human psychology, seems not to constitute a condition accountable in any relevant sense so as to license a correct application of blame to self-deceivers’ cases. We can’t possibly expect people to be so skilled to take a second-order stance over their cognitive strategy of the sorts Forrester describes – or self-deception would never occur. Still, interpreters’ dissatisfaction has a ground. The believers’ commitment to truth makes for the difference. It is in his motivational set80 to reach a reasonable conclusion: he wants his beliefs

79 M. Forrester, Self-Deception and Valuing Truth, cit., p. 42
80 A useful discussion on internal reasons and blame is in B. Williams, Internal Reasons and the Obscurity of Blame, in Making Sense of Humanity and Other Philosophical Papers,
to be rationally justifiable and he does not intend to believe what is false or unjustifiable. He wants to find explanations that will preserve the beliefs he favour, but he never fails to have the information at hand to realize that «his evasive activities could lead to his denial of the truth». The problem is that even once a proposition can look reasonable, we are in a position to choose whether to believe, disbelieve, or suspend judgement. ‘Given the fact that’ the self-deceiver never ceases to suspect that p might be false, it is rational from his own standpoint at least not to form a belief and suspending the judgment. ‘The formation of a belief under uncertain condition is never rational’. Other attitudes are recommended, such as doubting. Furthermore, that beliefs are sometimes formed under uncertainty does not entail that this is all the believing we live with. In particular, the subject matter of many cases of self-deception are highly verifiable through direct evidential proof (a cancer can be clinically tested, an unfaithful wife can be unmasked, and so on).

What we blame, I think, in our and others’ self-deception is that it is in principle among the self-deceiver’s capacities, qua rational agent, to compare his current method of dealing with evidence to his usual standards. This is the clue that normally alerts a self-deceiver to reconsider his psychological predicament and that normally rescues him from self-deception. On that occasion, he normally also realizes that the desire that p shaped his cognition in the way it did, and he gets an insight into the correct causal connection between his attitudes. So, the self-deceiver had many internal reasons not to end up forming a belief that it turns out to be false, in the first place, given the persistence of the suspect; and, secondly, the more he shows to be skilled in defending his propositions and evaluating his method of evaluation as exceptionally licensable given the circumstances, as described in section 4 and 5, the more we would want to expect him to have at least suspended the judgment, and taken one last, meta-cognitive stance about the suspicion that such an ad hoc epistemic evaluation gives rise to.

7. Conclusion.

I have argued that self-deception involves an intense rational activity and that it does not consist of a mere, immediate suspension of normal epistemic standards of a believer. I have also claimed that, although strong
intentionalism fails to give an unproblematic account of self-deception, and that deflationists tend to make self-deception appear in the guise of other, distinguishable forms of motivated irrationality, a weak intentionalism such as that defended in this essay can more plausibly account for self-deception. It fares better if coupled with some of the deflationist intuitions, e.g., that the agent fails to make an appropriate judgments over the causal drive of his epistemic behaviour.

I have also argued that blame applied to self-deception is grounded on the self-deceivers’ capacities qua reasoners interested in truth, as they intend themselves to be. I think a consequence of my account is that self-deception shows that the purpose of searching for truth is not enough in itself to guarantee that our search will actually lead us to truth. Impartiality seems then to be the crucial ingredient to gain truth, or at least to prevent us from straight believing a falsity, by recommending to us judgement-suspension where a suspect persists.82

82 This essay draws from researches begun in 2002 at the University of Florence on the occasion of my first degree dissertation and carried out at the University of Stirling (Scotland) during two semesters in 2004-2005. Special thanks to my supervisor Professor Alan Millar and Professor Duncan H. Pritchard for the bursary conceded as a full-time postgraduate student of the M.Litt. on «Knowledge & Minds» and for their unceasing help. I am grateful to Dr. Mike Wheeler (University of Stirling), Dr. Adrian Haddock (University of Stirling) and Dr. Peter Baumann (University of Aberdeen) for their comments on the manuscript. Thanks also to the University of Florence for the Ph.D. studentship previously granted and for the permission to spend one year at the University of Stirling. I am particularly in debt with Professor Alessandro Pagnini, his long-term studies in this area, his encouragement and supervision.