Self-deception and the selectivity problem

«Self-deception and the selectivity problem»

by Marko Jurjako

Source:
Balkan Journal of Philosophy (Balkan Journal of Philosophy), issue: 2 / 2013, pages: 151-162, on www.ceeol.com.

BALKAN JOURNAL OF PHILOSOPHY. Vol. 5, Issue 2, 2013

SELF-DECEPTION AND THE SELECTIVITY PROBLEM

Marko Jurjako

Faculty of arts and sciences at the University of Rijeka
Marohnićeva 4, 51000, Rijeka
Mob. 385 98 1866089
mjurjako@ffri.hr

Abstract:
In this article I discuss and evaluate the selectivity problem as a problem put forward by Bermúdez (1997, 2000) against anti-intentionalist accounts of self-deception. I argue that the selectivity problem can be raised even against intentionalist accounts, which reveals the too demanding constraint that the problem puts on the adequacy of a psychological explanation of action. Finally I try to accommodate the intuitions that support the cogency of the selectivity problem using the resources from the framework provided by an anti-intentionalist account of self-deception.

Key words: action explanation, folk-psychology, dispositions, epistemic virtue, Mele's deflationary account, self-deception, the selectivity problem.

1. Introduction

The phenomenon of self-deception has proven to be a very controversial philosophical subject whose paradigmatic cases are hard to characterize in a general manner. However, one can find a fairly neutral description of the phenomenon of self-deception that encompasses varieties of views on the Stanford Encyclopedia of Philosophy: "self-deception is the acquisition and maintenance of a belief (or, at least, the avowal of that belief) in the face of strong evidence to the contrary motivated by desires or emotions favoring the acquisition and retention of that belief" (Deweese-Boyd 2006).

Standard examples of self-deception that can be found in philosophical literature usually involve a person that is under the grip of emotion forming a self-deceptive belief about some state of affairs. For example, a mother is present at the court where her son is being tried for murder. She listens to all the evidence that the prosecutor is presenting and despite the persuasiveness of the evidence, which would be more than enough for an unbiased observer to convict her son, she stays convinced that her son is innocent. Intuitively, in this case we say that the mother has a self-deceived belief that her son is innocent because she is not responsive to the available evidence to the contrary and because the formation of her (self-deceptive) belief is caused in an irrational manner; it was caused by her attachment to the son and not by the available evidence.

1.1 Intentionalist and anti-intentionalist accounts of self-deception

There are two general groups of accounts that try to explain the phenomenon of self-deception. On the one hand, *intentionalists* think that self-deception essentially involves an intention on the part of the self-deceiver who actively (intentionally) deceives herself (e.g. Bermúdez 2000, Davidson 2004). On the other hand, *anti-intentionalists* think that self-deception can be explained without invoking the concept of intention in forming the self-deceived belief (see e.g. Johnston 1988; Lazar 1997, 1999; Mele 1998, 2001).

To illustrate differences between the two accounts let us consider Davidson's (2004) example:

> Carlos has a good reason to believe he will not pass the test for a driving license. He has failed the test twice before and his instructor has said discouraging things to him. On the other hand, he knows the examiner personally, and he has faith in his own charm. He is aware that the totality of the evidence points to failure. […] But the thought of failing the test once again is painful to Carlos (in fact the thought of failing at anything is particularly galling to Carlos). So he has a perfectly natural motive for believing he will not fail the test, that is, he has a motive for making it the case that he is a person who believes he will (probably) pass the test. (p. 209)

According to Davidson we should interpret Carlos' case as a case in which Carlos goes through a

process of practical reasoning that ends up with his forming an intention to form a belief that he will pass the test. Because of the uneasiness that Carlos would feel if he failed his test he forms a desire to avoid feeling this uneasiness. Together with the belief that this desire can be satisfied by forming a belief that he will not fail the test, Carlos intentionally leads himself to believing that he will not fail the test. This pattern of practical reasoning exhibited by Carlos is parallel to the reasoning that a person undergoes when intentionally doing something. For example, a construction worker wants to make concrete: she knows that in order to make concrete she must mix water, cement, and sand; therefore, she forms an *intention* to mix water, cement, and sand, and consequently she executes her intention into an (intentional) action.

However, intentionalist accounts traditionally face at least two problems, the so-called static and dynamic paradoxes (see Bermúdez 2000; Mele 2001).[1] These prima facie problems for intentionalist accounts gave a boost to anti-intentionalist proposals for explaining self-deception (cf. Bermúdez 2000, 309; Mele 2001; see also Bermúdez 1997). According to anti-intentionalists we do not have to postulate the presence of an intention in the mind of the self-deceiver in order to explain the phenomenon of self-deception. Anti-intentionalists claim that putative unintentional mechanisms of the mind are enough to explain garden-variety cases of self-deception. In this respect, the main strategy of the anti-intentionalists is to explain self-deception as a case of motivationally biased belief.

For example, anti-intentionalists would claim that Carlos does not intend to deceive himself; rather, they would claim that he has a strong desire not to fail the test, since that would cause him dire pain. Consequently, the desire biases his belief-forming process and causes him to believe that he is not going to fail the test. In the next subsection I will present what is currently the most influential anti-intentionalist account of self-deception and then in the following section I will present and discuss the objection to this account proposed by Bermúdez (1997, 2000)

1.2 Mele's anti-intentionalist account

According to the most influential contemporary anti-intentionalist, Alfred Mele (1997, 2001), there are four jointly sufficient, though not necessary, conditions that explain the acquirement of a belief through a process of self-deception:

1. The belief that p which S acquires is false.

2. S treats data relevant, or at least seemingly relevant, to the truth-value of p in a motivationally biased way.

3. This biased treatment is a non-deviant cause of S's acquiring the belief that p.

4. The body of data possessed by S at the time provides greater warrant for ~p than for p. (50-51)

The first condition is intuitive; Mele claims that "A person is, by definition, deceived in believing that p only if p is false" (ibid., 51), and that the same holds for self-deception.[2]

The second condition concerns the mechanisms through which people acquire self-deceived beliefs. Mele (ibid., 26-27) names four possible mechanisms. Negative misinterpretation is a phenomenon that happens when a person's desire that p causes her to misinterpret that some evidence does not count against p when the same evidence would be considered to count against p, if it were not for the desire that p. In our first example, it could be the case that the mother believes that the murder weapon with her son's fingerprints does not count as evidence for her son's being guilty, even though she would believe that it counts as evidence were it not for the desire that her son be innocent.

Positive misinterpretation is manifested when a person misinterprets some evidence as favoring her desired proposition when that evidence, in absence of the biasing desire, would easily be recognized as counting against the desired proposition. Mele (ibid., 27) gives an instructive example: Sid is in love with Roz, but Roz does not love Sid. However, Sid misinterprets her refusal of his affection as evidence that she really loves him (and pretends that she is 'hard to get'), and that she wants him to prove his real love for her.

Another psychological process is the selective focusing on evidence. The process in question is such that the mother's desire that her son be innocent can lead her to focus her attention on evidence that counts in favor of her son being innocent and that can lead her to fail to recognize the available

[1] To indicate briefly, the static paradox concerns the fact that intentionalists are seemingly committed to a claim that a self-deceiver has two contradictory beliefs (belief that p causes the belief that not p [see Davidson 2004, 208]). And the dynamic paradox reflects the seeming problem according to which the intentionalist needs to ascribe a self-defeating intention to a self-deceiver (i.e. the intention to deceive herself).

[2] However, not all authors would agree with this claim. For example Bermúdez (2000, 310–311, 312) argues that even in interpersonal deception the deceiver does not have to hold that the deceived belief has to be false, and by analogy the same should work for self-deception.

evidence that goes against the truth of her desired proposition.

The fourth mechanism that Mele presents is selective evidence-gathering. To illustrate this mechanism we can suppose that the mother is trying to guide her own investigation about her son's case. However, while investigating, the mother's desire that her son be innocent biases her into overlooking the evidence that goes against her desired proposition and leads her to find some less obvious and less convincing evidence that her son is innocent.

The third condition is added by Mele to secure the exclusion of cases of distorted belief-forming processes that are not cases of self-deception. For example, a mother may want to believe that her son is innocent and that desire can bias her search for evidence in favor of her desired belief. However, while trying to investigate the crime scene (or some scene more available to her) in order to find some evidence in her son's favor, she slips on the wet floor and falls on her head. After she wakes up, because of the trauma, she starts to believe that her son is innocent. In this case we can say that her desire to believe that her son is innocent caused her to believe that he actually is innocent. However, the causal chain from the desire that p to the belief that p is causally deviant in a way which prevents us from saying that this is a case of self-deception.

The fourth condition is meant to capture the often present characteristic of self-deception; that is, the failure of epistemic rationality. An epistemically rational person is someone who forms her beliefs in accordance with the best available evidence. However, a self-deceiver is a person who, despite the availability of good evidence to believe ~p, starts to believe that p because she has a biasing desire that p be true. So, a self-deceiver is epistemically irrational because she forms her beliefs on the basis of her desires and not on the basis of the available evidence. However, unlike Davidson (2004), Mele does not take this condition to be necessary for entering into the process of self-deception.[3]

In order to make his account more empirically adequate, Mele (1997, 2001) relies on an empirical model of everyday hypothesis testing, developed by the psychologists Friedrich (1993), and Trope & Liberman (1996). The idea behind the FTL model[4] is succinctly summarized by Bermúdez (see also Mele 2012):

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; similarly for rejection. Hypotheses which have a high acceptance threshold will be more rigorously tested and evaluated than those which have a low acceptance threshold (Bermúdez 2000, 316).

In garden-variety cases of self-deception Mele proposes that the structure of everyday hypothesis testing will enable one to enter into a process of self-deception. This idea can be illustrated with an example. A mother who desires that her son is innocent would experience more painfully the thought that her son is guilty than that he is innocent. These emotions, and the desire that her son is innocent, structure her expected subjective cost of believing a hypothesis in a way that makes it easier for her to falsely believe that her son is innocent, than to falsely believe that her son is not innocent. Because of this subjective cost, her acceptance threshold for believing that her son is innocent will be much lower than the acceptance threshold for believing that her son is guilty. According to Mele, the desire that the son be innocent and the low threshold associated with the desired state of affairs is what explains the mother's acquiring the self-deceived belief that her son is innocent.

2. Anti-intentionalism and the *selectivity problem*

Bermúdez (1997, 2000) has put forward a problem that threatens to undermine anti-intentionalist accounts in favor of intentionalist models. In this section I will present his objection. Here is how Bermúdez formulates the problem:

Any explanation of a given instance of self-deception will need to explain why motivational bias

[3] In a recent article Mele (2012) introduces two amendments to his jointly sufficient conditions for acquiring self-deceptive belief. These are: condition 5: "S consciously believes at the time that there is a significant chance that ~p"; and condition 6: "S's acquiring the belief that p is a product of "reflective, critical reasoning," and S is wrong in regarding that reasoning as properly directed" (12). Condition 5 is meant to accommodate the intuition that there is some cognitive tension (conflict) present in cases of self-deception. While condition 6 captures the intuition that self-deception requires mistaken higher-order thought about the causes of one's own cognitive activities (see Mele 2012, 10–12). I am mentioning these two amendments in a footnote because they do not play any significant role in what follows in the present article.

[4] FTL stands for Friedrich, Trope and Liberman, psychologists who developed the model of everyday hypothesis testing.

occurred in *that* particular situation. But the desire that p should be the case is insufficient to motivate cognitive bias in favour of the belief that p. There are all sorts of situations in which, however strongly we desire it to be the case that p, we are not in any way biased in favour of the belief that p. How are we to distinguish these from situations in which our desires result in motivational bias? I will call this the selectivity problem (Bermúdez 2000, 317).

According to Bermúdez, the conditions enumerated by Mele for having a case of self-deception are insufficient because they cannot explain why the desire that p be the case in one situation will cause us to believe that p is the case, while in another situation it will not. Bermúdez (1997, 108) claims that Mele illegitimately draws "analogy between familiar examples of unintentional cold bias and motivationally hot bias." The difference between instances of cold bias and self-deception is that cold bias is *nonselective*, which means that regardless of its subject matter it gets activated, while the process of self-deceptive belief formation is subject dependent (ibid.).

Mele (2001, 63-65) tries to answer the selectivity problem by using the FTL model for hypothesis testing in order to explain what makes the difference between two situations in which a desire with the same content has different effects on the formation of the belief. He gives an example of Gordon who is a CIA agent that has been accused of treason. Gordon's parents and all of his colleagues want him to be innocent and they all have access to the same information regarding Gordon's case. However, based on the available evidence, Gordon's parents come to the conclusion that he is innocent, while his colleagues, on the basis of the same evidence, conclude that he is guilty.

According to Mele, what explains that the parents and Gordon's colleagues come to believe different things, despite having the same desire and access to the same evidence, is the fact that their acceptance/rejection thresholds for the hypothesis that Gordon is innocent are different. For Gordon's parents it is less costly to falsely believe that he is innocent, than to falsely believe that he is guilty. While for his colleagues it is the other way around; since their lives depend on the verdict, it is much costlier for them to falsely believe that Gordon is innocent than to falsely believe that he is guilty. Accordingly, the difference in associated costs causes parents to have a low acceptance threshold for the hypothesis that their son is innocent, while his colleagues have high thresholds for the acceptance of the same hypothesis.

However, Bermúdez (2000, 317–318) asserts that the selectivity problem reappears again even if we introduce the cost/benefit analysis as Mele does. Bermúdez's claim is that "there are many hypotheses for which my motivational set dictates a low acceptance and high rejection threshold and for which the evidence available to me is marginal enough to make self-deception possible. But I self-deceivingly come to believe only a small proportion of them. Why those and not others?" (2000, 318).

Bermúdez's answer is that what is missing and what solves the selectivity problem is the intention "on the part of the self-deceiver to bring it about that he acquires the belief that p" (ibid.). This is why, according to Bermúdez, intentionalist accounts (in some form) must be right, since they have a ready answer to the selectivity problem that anti-intentionalists supposedly do not have.

In the next section I will try to show that, if genuine, the selectivity problem can also be raised against intentionalist accounts. After that I will examine one possible solution to the selectivity problem proposed in (Pedrini 2010).

3. The *selectivity problem* for intentionalists

If the selectivity problem is genuine, then introducing intentions into the account will not solve the problem. As Mele (2001, 65-66) already noticed, even if intentions were necessary for accounting for an episode of self-deception, intentionalists will have a selectivity problem of their own. Let us suppose that what explains the difference between situations in which agent A with a strong desire that p self-deceivingly believes that p and situations in which she does not, is the fact that in one situation she had an intention to bring it about that she believes p and in another she had no such intention. Nevertheless, it is clear that there are many cases in which we intend to do something and fail to accomplish it. I can intend to become a basketball player, and fail to become one; also, I can intend to raise a glass that is on the table and fail to do that, just like I can intend to hit the bulls-eye and not manage to do that. So, having an intention to bring oneself to believe something does not seem to be sufficient for entering into a process of self-deception since nothing in the nature of intention guarantees that the intention will be effective. Hence, the question of selectivity of self-deception would arise for intentionalists as well; they seem to be faced with the question of what is the difference between situations in which agent A has the intention to be self-deceived and manages to self-deceive herself and

situations in which she does not succeed in executing that intention.

Moreover, a stronger point can be made against the intentionalist who thinks that the selectivity problem can be resolved by introducing the relevant intention. In his (1997, 108; 2000 317) Bermúdez put the following desideratum on the proper explanation of self-deception: "Any explanation of a given instance of self-deception will need to explain why motivational bias occurred in *that* particular situation." If we generalize Bermúdez's proposal, we can say that at the most general level, the selectivity problem boils down to the question: why some people deceive themselves and others do not?[5] From this standpoint, even if, for the sake of argument, we grant that intentionalist accounts are true, we still would not have a proper explanation why in a particular case an episode of self-deception occurred. Namely, Bermúdez supposes that introducing an intention to acquire a belief will explain why in *this* particular situation an episode of self-deception occurred. However, an intention by itself cannot explain an occurrence of self-deception. To rephrase Bermúdez (2000): Any explanation of a given instance of self-deception will need to explain why *intention* was formed in that particular situation.

To see why let us consider a more general example of an instance of action explanation. In folk-psychological explanations of action we usually explain why someone did something by invoking the reasons for which that person acted. For example, we say that Jennifer went to a store because she wanted some candy, and that she believed that by going to the store she could buy some candy. Since we suppose that Jennifer's action was intentional we ascribe to Jennifer an intention to go to the store. However, ascribing an intention to Jennifer does not explain *why* she went to the store; the reasons (i.e. desires and beliefs) explain why she formed that intention and performed that action. A similar point can be made in the case of self-deception. Introducing an intention to self-deceive will not explain why in this particular situation an instance of self-deception occurred. We need to know the reasons for which the intention is formed, and, in the folk-psychological sense, this will include the desire and a means-end belief.[6]

So, in order to explain why in this particular instance self-deception occurred, we need to invoke a desire and a belief. But, now, we can ask why in this particular situation a desire that p be the case caused an intention to believe that p is the case? As Bermúdez noted, we have all kinds of desires that, nevertheless how strong, do not cause us to believe that p is the case; similarly we can say that we have different strong desires to believe that p be the case (or that we believe that p is the case), that nevertheless do not cause an intention to believe that p. So, in this way we can raise the selectivity problem against the intentionalist account. Namely, we can raise the question why in this *particular* situation desire that p be the case (or to believe that p) caused an intention to believe that p is the case since, according to Bermúdez, in all kinds of situations, no matter how strongly we desire that p be the case it does not cause us to believe that p is the case.

Someone could argue on behalf of Bermúdez that the intentionalist account of self-deception explains the occurrence of an episode of self-deception and solves the selectivity problem as a complex whole that includes intentions and reasons for the intention. She can say that the selectivity problem only indicates the possibility of a situation in which we have two persons (we can call them A and B) with equally strong desires that p be the case and equal thresholds for accepting and rejecting hypothesis, of which only one person self-deceivingly starts to believe that p is the case. And the intentionalist supposes that what explains the difference between A and B is the formation of the intention to believe that p on the part of the individual that self-deceives (e.g. A).[7]

However, this response could only work, as a solution of the selectivity problem, if we suppose two things (not necessarily as a conjunction) about the situation[8] in which we have A and B. (i) We can suppose that intentions to self-deceive are caused by sheer chance by the reasons that precede the inten-

[5] I am very grateful to Patrizia Pedrini for giving me an idea how the following case can be used against the intentionalist's solution of the selectivity problem.

[6] See Bermúdez (1997, 108) where he writes: "the intentionalist holds that the subject must intend to cause himself to believe that *p* by biasing his cognitive processes because (a) he desires to believe that *p* and (b) he believes that the best way to achieve this is to bias his cognitive processes in the ways that Mele discussed."

[7] One could object that the example with two persons is not adequate since Bermúdez explicitly writes that the "selectivity problem is not a problem of how two people in similar situations can acquire different beliefs" (2000, 317). However the example would work even if we suppose that there is one person with an identical psychological makeup at different moments in time.

[8] This situation, by supposition, involves complete psychological (functional) identity between A and B in moments before A forms an intention to acquire some belief.

tion formation. Or (ii) we can suppose that intentions to self-deceive are an effect of a conscious decision.

Response (i) does not sound very plausible, since we would need to suppose that what precedes formation of an intention does not influence in any substantive way the intention to form a self-deceiving belief. This would go against our folk-psychological conception of ourselves, which include law-like generalizations that figure in action-explanations (see e.g. Churchland 1981).

If we suppose (ii) then we could explain the difference between A and B in a plausible way, since we could say that A (unlike B) deliberated about the issue at hand and decided that she will intentionally deceive herself into believing that p is the case. As a consequence of that decision A could then engage in performing the necessary steps to cause herself to believe that p. However, response (ii) is not plausible as an intentionalist account of self-deception because it could not encompass and explain garden-variety cases of self-deception. Namely, it overly intellectualizes the core cases of self-deception since it supposes that they have the form similar to Pascal's wager and this is something that even Bermúdez (2001, 318) recognizes as not being vindicated by his argument. Furthermore, such overly intellectualized cases could even be incorporated in anti-intentionalist accounts since they do not suppose that a person can never decide to acquire a belief and then to undertake the necessary steps to accomplish that goal. However, the crucial thing that divides the two accounts of self-deception is exactly the explanation of garden-variety cases of self-deception (see Mele 1997, 99).

Hence, it seems that intentionalist accounts do not present a proper explanation of garden-variety cases of self-deception. Since, if genuine, the selectivity problem can be raised even against them.

4. Value-based solution to the selectivity problem

More recently Pedrini (2010) has offered a possible solution to the selectivity problem by supposing that what makes the difference between a self-deceiver and a person who resists self-deception is the *embodiment* of epistemic values that the latter exhibits. To embody an epistemic value is to be disposed to behave in certain ways. Here is an example from ethics:

[i]f *S* embodies the value of generosity, to the extent that generosity can be attributed to *S* as one of her traits, *S* is disposed to a wide range of behaviors that we define as generous: perhaps she offers her friends a dinner to celebrate their birthdays, or she will help others in need even if they are strangers, etc. (Pedrini 2010, 133.)

Similarly, Pedrini claims that a person who *embodies* epistemic values, such as a value of truth, epistemic accuracy, completeness and impartiality (ibid.) will be disposed to behave in such a way that will preclude that person from entering into a state of self-deception. Thus, according to this account, a person can have a desire that p, a desire that could cause her to lower the acceptance threshold for the hypothesis that p, but she will not be disposed to self-deceive if she really embodies the epistemic values protective against self-deception. For example, if a mother has a desire that her son is innocent and has a low acceptance threshold for this hypothesis, if she embodies the relevant epistemic values, which constitute her dispositions to behave in an epistemically virtuous way, she will override her biasing desire and start to evaluate the evidence in order to find out the truth, no matter what are the possible costs associated with finding new and possibly devastating evidence for her son's case.

Pedrini's account is an interesting *prima facie* solution to the selectivity problem. It introduces the evaluative element by supposing that what explains the selectivity of self-deception is the presence or absence of epistemic values to which the agent is committed. Moreover, it seems that Pedrini's account can improve and amend Mele's use of the FTL model, since now we can explain what, besides desires, determines the rejection/acceptability thresholds for hypotheses. In an epistemically virtuous agent, her epistemic values determine the relevant threshold, and protect from its being affected and modified by desires. So, using Pedrini's account we can explain the difference between agents who deceive themselves and ones who do not by appealing to their evaluative structure.

However, there is an objection that can be pressed against this latter account. Pedrini (2010, 134-135) examines one possible objection. One might say that epistemic values cannot be enough to secure one from exhibiting an episode of self-deception because one can be weak-willed and in the presence of a desire that p can become akratic and unable to employ their best epistemic strategies. That is, in that case, a person would, against her best epistemic judgment about how to treat the evidence, be unable to act in accordance with that judgment, and consequently the desire that p would cause her to treat available evidence in a motivationally biased way. So in this kind of case (epistemic *akrasia*) embodiment of epistemic virtues apparently would not secure one from forming self-deceptive beliefs, and

furthermore, would not resolve the selectivity problem.

Pedrini tries to show that the case of akrasia is not a counterexample to her account because her account of embodied values is *designed* to rule out cases of epistemic *akrasia*[9] (Pedrini ibid, 134). She explains her idea by introducing the distinction between "something being judged valuable" and "valuing something" (ibid.). For example, one may say that helping others is good, but then never exhibit the helping behavior when in a situation where *helping* behavior is appropriate. For that person we would say that she does not really *value* helping others. On the other hand, if someone values something, then if the right circumstances occur, she will be *disposed* to act in accordance with her judgment. According to Pedrini (ibid.), to *really value* something is to *embody* that value and if we really embody a value then it seems that we cannot fail to exhibit that value when the circumstances are appropriate.

However, the introduction of the distinction between 'judging valuable' and 'valuing' does not get to the core of the problem of the possibility of epistemic *akrasia*. In order to evaluate the plausibility of the latter counterexample (and its supposed rebuttal) we need to restrict our attention only to cases of 'valuing', where this attitude can be construed as being manifested in a *sincere* judgment that something is valuable.[10] So what is 'valuing'? According to Pedrini (2010) to value something is to embody that value, and to embody a value is to be disposed to act in accordance with the corresponding sincere value judgment when the circumstances are appropriate. Therefore, it seems that what plays a crucial role in explaining valuing is the notion of a *disposition*, since disposition explains what it means to *embody* a value. The notion of a disposition involves all kinds of intricacies that have not yet been philosophically resolved; however there are a few commonplaces (platitudes) concerning dispositions that are relevant for the present article. It is widely recognized that dispositions involve some notion of conditionality that connects their *stimulus* and *manifestation* conditions.[11] For example, let us take into consideration the notion of *fragility*. Typically, the fragility of an object manifests itself in the object's tendency to break under suitable conditions. However, to ascribe fragility to an object it is not necessary that its typical manifestation always occurs; a fragile object will still be fragile even if it never manifests that property (e.g. it never breaks). Dispositional properties will be manifested only if certain conditions are satisfied, that is, only if certain antecedent or *stimulus* conditions are fulfilled (cf. Mumford 1998, 6). Since veridical ascription of dispositions does not require occurrence of typical manifestation, a thing can have a dispositional property that is never manifested. For example, an object can be fragile, without ever breaking because stimulus conditions for its becoming broken never occur. Also it is possible that even if stimulus conditions are present the circumstances in which manifestation usually occurs do not have to obtain because some interfering factors are also present. For example, when emerged into the water (stimulus condition), sugar will typically dissolve (manifestation of solubility), but if the water were already saturated, then sugar's solubility would not be manifested.

These same features of dispositions when translated into the talk about judgments and value take the following form. For example, a plausible formulation of value judgment internalism, given by Michael Smith, takes the following form: if a person judges that it is right to do A, then if rational, she will do A (Smith 1994, 62). According to this view, if someone does not behave in accordance with her sincere value judgment, it does not mean that she does not embody the value expressed by that judgment, it only means that she is exhibiting some kind of irrationality (perhaps she is depressed or weak-willed, etc. (ibid.)); that is, it just means that the disposition to act in a way that corresponds to one's sincere value judgment does not get manifested.[12] Now the question can be posed: does the notion of an epistemically virtuous agent (or an agent that embodies epistemic value) play a role similar to *stimulus* conditions (that encompass the rationality

[9] When I use the expression 'epistemic *akrasia*' I intend it in the sense in which it is used in theory of action with the obvious limitation, since in the epistemic case, the action will be concerned with theoretical investigation whose final product is a belief (and I do not presuppose that the belief formation itself needs to be intentional). Accordingly, epistemic *akrasia* will concern the failure of employing epistemic strategies that one judges to be the best (see Pedrini 2010, 134–135).

[10] When I say sincere judgment I mean to include the idea that the agent not only believes of herself that her value judgment is sincere (where the latter belief could be mistaken), but that the value judgment really (actually) manifests what she really (actually) *values*.

[11] What kind of conditionality will the notion of a disposition involve and whether it will always be a counterfactual, is a controversial philosophical issue. For a discussion see e.g. (Bird 2007, section 2.2); Mumford 1998, chapters 3, 4).

[12] Still in other words, we can say that the disposition gets defeated, but not erased.

condition) or the role similar to the dispositional property itself that manifests certain properties in the right *stimulus* conditions?[13]

If we equate the embodiment of epistemic values with the disposition to exhibit epistemically virtuous action then the problem of epistemic *akrasia* lurks back in. Namely, if embodiment of epistemic values involves dispositions to act in certain ways, then the ascription of the latter disposition(s) can be veridical without manifesting its consequence(s) in any particular case. For example, if we take the latter line, it can still be the case that a person self-deceives that p is the case even though she still has the disposition to behave in an epistemically virtuous way. To take the standard philosophical platitude, the disposition would be defeated, but not erased, because the desire that p be the case is so overwhelming in this particular situation that it manages to *silence* (defeat) the otherwise epistemically stabile character. Thus if we take the latter reading as something that Pedrini proposes then the case of epistemic *akrasia* becomes a problem for her account once again.

Hence, it seems that in order to avoid the former consequence we should equate epistemic value embodiment with *stimulus* conditions or, in Michael Smith's sense, with the rationality condition. If we take this line of thought then it seems that we can exclude cases of epistemic *akrasia*. Namely, just like in the practical case, where a person who does not act in accordance with her sincere practical judgment exhibits a case of practical irrationality, we can say that when a person self-deceives she exhibits behavior (or thought) that is *ipso facto* not epistemically virtuous.

However, even if plausible to a certain degree, the value-based solution of the selectivity problem is not without its problems. For example, we can ask whether the embodiment of epistemic values come in degrees or is an absolute characteristic; and connected to the latter issue, whether we should take the embodiment of values to be contextual, so that one can fully embody one and the same value in certain life circumstances and fail to embody it in another, or not.[14] If we suppose that embodiment comes in degrees then it would seem that the problem of *akrasia* could again become salient because there is no *obvious* boundary that would delimit degrees of embodiment that would exclude cases of *systematic akrasia. Prima facie*, it seems that we could only be certain that full (absolute) embodiment of values prevents cases of systematic *akrasia* to occur. So, *prima facie*, it seems that a person who is not fully epistemically virtuous could in certain circumstances always be overcome with a desire that p and consequently be induced in believing that p; and there does not seem to be any obvious reason to suppose that this akratic self-deception cannot be systematic.

We can avoid this potential problem by supposing that the embodiment of epistemic values should be taken in an absolute (non-degree) sense. However, in that case one could object that the solution of the selectivity problem in terms of embodiment of epistemic value loses its explanatory power. The idea is that one could admit that on the general level non-embodiment of epistemic values explains why self-deception occurs; usually people self-deceive because they are not epistemically virtuous. However, one could still argue that proper non-embodiment of epistemic values does not explain Bermúdez's original formulation of the selectivity problem, that is, why self-deception "occurred in *that* particular situation" (Bermúdez 2000, 317). Since most people do not exhibit stabile epistemic dispositions (they are not epistemically virtuous [see e.g. Samuels and Stich 2004]), the fact that they do not fully embody epistemic values would not have much explanatory power to explain why self-deception occurred in some, but not in other particular cases. In this case, someone could say that the selectivity problem becomes acute again even for the account that introduces embodiment of epistemic values. At this point one might again argue that two persons can have the same desire that p, be at the same (non-perfect) level of epistemic value embodiment, and have the same rejection/acceptance thresholds and still differ with regard to the exhibition of self-deception (one self-deceives, while the other does not). Consequently, that person might argue again that what is needed is the intention on the part of the agent in order to explain her exhibiting self-deceiving belief. However, as I argued earlier, I do not believe that this strategy actually works as a solution of the selectivity problem.

The latter remarks about the value-based account are not meant to prove the falsity of the account; however, they do point out some of the gaps

[13] Of course an epistemically virtuous agent will embody a whole set of epistemic values and correspondingly will be disposed to exhibit a whole range of subdispositions that comprise epistemic virtues (such as evidence gathering, evidence evaluation, sound reasoning, etc.).

[14] For example, a successful scientist may embody all the relevant epistemic values in the context of her professional work, but be completely unresponsive to them when issues concerning her private life become salient. Does such a person fully embody relevant epistemic values in one context but not in the other, or does she overall embody them to a certain degree, etc.?

or misgivings of the account. Pedrini (2010, 135) also claims that her account is not a final word on the matter but that it provides a framework in which the selectivity problem could be solved. In section 5 I will try to show why I find the selectivity problem dubious, and how the intuitions that support its genuineness can be explained away by using the framework of Pedrini's value-based account.

5. Dissolution of the *selectivity problem*

At this moment one might start to despair that there is no solution of the selectivity problem, since the same problem can be raised even against intentionalist accounts, and the other proposed account (value-based account) does not seem to be adequate. However, I believe that the discussion in section 3 enables us to see the mootness of the *selectivity problem* as a genuine problem that sophisticated accounts (such as Mele's 2001) of self-deception face. In section 3 I stated that, if genuine, the selectivity problem can be directed towards intentionalist accounts[15], because we can always raise the question of why in this particular situation a desire that p be the case (or to believe that p is the case) has caused one to acquire an intention to form a belief that p is the case.

However, I believe that an intentionalist may properly solve the selectivity problem (as posed in section 3) simply by saying that that is just how the mechanism works, or at least how our folk-psychological explanation, in simple cases, depicts the workings of the mechanism that produces intentional action. In normal cases, we have a strong desire that p and a belief that if I do q I will accomplish p, which causes us to decide to form an intention to make q true. Similarly in the case in which we adopt an intention to self-deceive that p is the case. I think that from a folk-psychological perspective the selectivity problem puts too demanding a requirement on the adequacy of an action-explanation; if we take it as a genuine problem then it would be baffling how even in everyday cases an intention gets formed on the basis of the reasons that precede it.

However, if we grant this answer to the intentionalist then by parity of reasoning we have to grant it to the anti-intentionalist as well. The only difference will consist in the fact that the anti-intentionalist will claim that it is not necessary to postulate an intention to acquire a belief in order to account for garden-variety cases of self-deception.

[15] At least against those accounts that have the form displayed in (Bermúdez 1997, 108).

At this point the question that arises is why we feel the intuitive cogency of the selectivity problem as proposed by Bermúdez (1997, 2000)? The intuitive pull of the problem might be invited by Mele's assertion that his conditions for self-deception are *conceptually* sufficient (see e.g. his [2012, 2]). This idea of conceptual sufficiency of Mele's conditions should be read as giving a proto-analysis (ibid.) of the *self-deception* concept, and presumably its application conditions. So, one might wonder and doubt whether it is really *conceptually sufficient* that when Mele's conditions are satisfied then somehow by conceptual necessity it follows that we have a case of self-deception.

However, it seems to me that, regardless of what is meant by *conceptual sufficiency* in this context, when it comes down to explaining actual cases of self-deception the sufficiency of the thresholds for hypotheses acceptance and rejection (modified by the desire that something be the case) must include a *ceteris paribus* clause since psychological explanation is a paradigmatic species of explanation that demands *ceteris paribus* laws (see e.g. Churchland 1981, Davidson 2001, Fodor 1991), and if I understand Mele (2001, 2012) correctly, in giving a deflationary account of self-deception he intends it to be psychologically adequate. Hence, the question that now becomes pertinent is in what sense desire's lowering the acceptance threshold of a particular hypothesis is not sufficient for self-deception?

Logical sufficiency (in the sense that the threshold for a hypothesis entails self-deceptive formation of a belief in that hypothesis) presumably cannot be adequate, since an empirical phenomenon does not presuppose logical necessity. Thus, just because we can imagine that a motivationally modified hypothesis testing mechanism (that satisfies Mele's other conditions) does not *necessarily* lead to self-deception is not an argument against its empirical sufficiency. On the other hand, if we take it that the notion of sufficiency is causal then this notion of causality cannot simply involve *nomic necessity* since self-deception depends on contingencies of human nature (its less than perfect *epistemic nature*).[16] Thus we need to take into account the notion of *ceteris paribus* condition. But when we consider the selectivity problem in this light then the problem itself loses its weight. The question 'why one self-deceives in one situation and not in others' (Ber-

[16] For example, when Bermúdez states "[e]ven if one desires both that *p* be true and that one come to believe that *p* it is not *inevitable* that one will form the belief that *p*" (2000, 318, my emphasis) it is possible that he has this strong reading of sufficiency in mind.

múdez 2000, 318) in a sense dissolves because the anti-intentionalist may argue that cases where self-deception does not occur even though the acceptance/rejection thresholds are appropriately motivationally modified comprise cases in which *cetera* are not *paria*,[17] that is those are cases in which normalcy conditions for the occurrence of self-deception do not obtain.[18]

So, my claim is that from the perspective of a (folk) psychological explanation there is no problem for a sophisticated anti-intentionalist to explain why in certain situations a desire can be sufficient to induce a self-deceiving belief and in others will not be sufficient. The dissolution of the selectivity problem, according to this view, consists in taking seriously into account the *ceteris paribus* conditions that psychological explanations contain. Now, this invoking of *ceteris paribus* conditions might seem as a sort of magic bullet. In order to dispel these kinds of doubts, in the next section I will sketch a model of the conditions in which self-deception can be reasonably expected to occur.

5.1 Amending the value-based account of self-deception

Earlier, in section 4, I argued that Pedrini's (2010) value-based account could involve two notions of epistemic virtue: epistemic virtue can be fully (absolutely) embodied or it can be embodied to a certain degree. On the one hand, the problem with the notion of full embodiment was that it was not very explanatory in the context of the selectivity problem since people are usually very far from embodying epistemic virtues. On the other hand, the seeming problem with the graded notion of embodied epistemic virtues was its inability to determine when a real cognitive agent is secured from self-deception or will undergo a biased treatment of evidence caused by a desire that something be the case, since ordinary cognitive agents regularly fall short of embodying epistemic virtue (usually to a varying degree).

However the above remark concerning the graded conception of epistemic value embodiment is only a problem if we think that desires are not sufficient to bias one's cognitive abilities, other things being equal. But if we take desires to be sufficient, other things being equal, then grades of distance of embodiment of epistemic value from the ideal can figure in the sufficiency clause as a part of the normalcy condition that characterizes the background of desires' (that cause self-deception) *ceteris paribus* clause.[19]

Let me try to develop the latter thought a bit. As I interpret it, according to anti-intentionalists (of Mele's deflationary kind) a desire that p be the case is sufficient, *ceteris paribus*, to bias one's cognitive abilities and consequently to produce self-deceptive belief. *Ceteris paribus* condition involves the background conditions that constitute the *normal* circumstances in which desire causes self-deception. In our case, normal conditions that allow self-deception to occur, among other things, involve the condition that self-deceivers embody epistemic virtues to some degree.

So, how can this vague idea of a degree of embodiment of epistemic values play any explanatory role in this account? The answer I have in mind is provided in the following rough sketch: when relevant conditions are satisfied and there is a strong desire that something be the case, then, normally people (who have that desire) self-deceive.[20] That people normally self-deceive means that the statistical probability of self-deception is sufficiently high given the aforementioned conditions.[21] So where do degrees of epistemic value embodiment enter the picture? They enter at the level of background normal circumstances; the supposition is that what determines the statistical probability of forming the self-deceptive belief that p given the desire that p is the distance of the behavior of the actual epistemic agent from its corresponding ideal counterpart.

[17] Despite the fact that Bermúdez claims the opposite. See (Bermúdez 2000, 317) where he claims that: "[…] possessing a desire that *p* be true is not sufficient to generate cognitive bias, even if all other things are equal […]."

[18] This also refutes Pedrini's contention that desires are not sufficient to cause motivationally biased treatment of evidence that leads to self-deception (see Pedrini 2010, 134). Since conditions in which desires become sufficient to cause biased treatment of evidence that lead to self-deception exclude conditions in which an agent fully embodies epistemic virtues (i.e. *cetera* are not *paria*).

[19] *Ceteris paribus* clauses (and scientific laws that use them) are notoriously hard to adequately characterize. For a taxonomy of different senses of *ceteris paribus* sentences see Schurz (2002). However, it seems that recently a *prima facie* consensus (or a promising optimism) emerged that *ceteris paribus* clauses may be characterized by using so called normality accounts (see e.g. Reutlinger, Schurz and Hüttemann 2011).

[20] Or more specifically people exhibit confirmation bias that leads them to self-deceive, but not to complicate things I will just write as if desire in normal situations lead directly to self-deception.

[21] This is called the statistical probability consequence thesis. The thesis is somewhat controversial, but see Schurz (2001) for a forceful defense.

To make this idea more precise let us represent an actual cognitive agent and her corresponding ideally epistemically rational counterpart with functions Ψ and Φ, respectively. Cognitive agents can be described as mappings between input world states (or states as perceived by the agent) and output mental (belief) states, e.g. mapping Ψ takes evidence (world states) and outputs hypothesis (belief states) such that $\Psi(e_i)= h$, where i is an index of some evidence set.

We can define a distance between the real and the ideal cognitive agent in the following way: let d be a distance function that takes Ψ and Φ as its arguments. Also let ρ be a function from the set of cognitive agents to real numbers between zero and one such that $\rho_i(\Psi)=1$ if agent Ψ behaves in a fully epistemically rational way (fully embodies epistemic virtue) and $\rho_i(\Psi)=0$ if Ψ behaves in a completely irrational way from the epistemic perspective (does not embody epistemic virtue at all).[22] In other words, the idea is that ρ gives a degree to which agents rationally treat/use/interpret evidence and form/evaluate hypotheses and it tries to capture the rationality of an agent's behavior across different situations, where index i equals some natural number and signifies the difference in contexts in which a cognitive agent may exhibit (or not) its epistemic virtues. As a tentative and provisional formalization we can take the following equation to determine the distance function between the actual cognitive agent Ψ and its ideal counterpart Φ[23]:

$$d(\Psi,\Phi) = \frac{1}{N}\sum_{i=1}^{N}(\rho_i(\Phi) - \rho_i(\Psi))^2$$

Since Φ fully embodies epistemic virtues ρ_i will always assign her 1, but $\rho_i(\Psi)$, given that Ψ is a real, cognitively (and otherwise) limited, agent can take any value between zero and one. Hence, if we take only the extreme possible values, then across situations, if $\rho_i(\Psi)=1$ then $d(\Psi,\Phi)$ will equal zero. Which means that there is no difference between the actual agent and its epistemically rational counterpart in that particular situation; or we can say that Ψ fully embodies and manifests in a particular situation the epistemic virtue. If $\rho_i(\Psi)=0$ then $d(\Psi,\Phi)$ will equal one which would mean that Ψ is at a complete distance (according to this measure) from its ideal counterpart and accordingly does not embody epistemic value.[24] Thus in general, $d(\Psi,\Phi)$ will equal the average difference between the epistemic behavior of Ψ and Φ, across different situations in which cognitive agents might find themselves.

Using the above formalization we can express the idea that the distance between the ideal epistemic agent and the real agent determines the statistical probability that the agent will be prone to epistemically irrational behavior, and the greater the distance between Ψ and Φ the more likely it is that the a desire that p be the case will influenceΨ's belief forming processes. Hence, the *appropriate* values of the distance ($d(\Psi,\Phi)$) measure constitute the background normal conditions that need to obtain in order for the desire to be able to influence (cause) one's self-deceiving belief formation. I wrote appropriate values of d since, if e.g. $0.09 \geq d(\Psi,\Phi) \geq 0$ then presumably desire that p be the case will have negligible or no influence on the Ψ's belief forming processes. However, since we suppose that self-deception is a real phenomenon then there must be appropriate distance between Ψ and Φ (relative to some standards of rationality or virtues that an ideally rational being embodies) which constitutes the background conditions that allow the phenomenon to occur given that the desire that something be the case is present (and suitably intense or strong), that is, that the statistical probability of self-deception (given that a certain desire occurs) is sufficiently high.[25] So, the basic idea is that the greater the distance between the actual and ideal cognitive agent the more probable it will be that the actual cognitive agent self-deceives that p given her desire that p be the case.

From the perspective of the ideas developed in this article Bermúdez's (2000) claim that "possessing a desire that *p* be true is not sufficient to generate cognitive bias, even if all other things are equal (…)" (p. 317) is not adequately supported. The dissolution of the selectivity problem, in the framework of folk-psychology, invokes the idea that the desire that p be the case causes motivational bias, *ceteris paribus*; where *ceteris paribus* conditions include an appropriate distance degree between the real cognitive agent and her ideal counterpart or equivalently that there is a certain average difference between the

[22] What determines the standards of epistemic rationality or epistemic virtue is a matter for epistemology and philosophy of science to discuss and settle. Here I will remain silent regarding the determination of standards of epistemic rationality.

[23] One can immediately see that the equation is an instance of the Brier Score.

[24] Or we can say that she embodies the epistemic value to a degree 0.

[25] And ideally d should be empirically measurable. One of the conditions for this depends on our being able to agree on which epistemic standards should be used as standards of evaluation of cognitive abilities and generally epistemic behavior of individual agent.

real and the ideal that constitutes the *normal conditions* that *allow* or create circumstances in which a desire can cause biased treatment of evidence and consequently self-deceptive belief formation.

6. Conclusion

In this paper my aim has been to discuss and examine the selectivity problem as proposed by Bermúdez (1997, 2000). Bermúdez presented the *selectivity problem* as a problem that counts against anti-intentionalist accounts of self-deception. Bermúdez also argued that intentionalists have a ready-made solution to the selectivity problem and consequently that this problem counts in favor of intentionalists' accounts of self-deception.

I argued that the selectivity problem can be raised even against intentionalist accounts. Furthermore, I argued that the fact that a selectivity problem can be raised against intentionalist accounts indicates the way in which the problem could be dissolved. The latter fact shows that the insistence on the cogency of the selectivity problem puts a too demanding requirement on what constitutes a sound (folk) psychological explanation. Therefore, in the final sections I set out to analyze the intuitions that gave a convincing tone to the selectivity problem and tried to accommodate them using resources that do not go beyond the commitments of a sophisticated anti-intentionalist account.

References

Bermúdez, J., L. (1997). "Defending intentionalist accounts of self-deception." *Behavioral and Brain Sciences*, 20: pp. 107–108.

Bermúdez, J., L. (2000). "Self-Deception, Intentions, and Contradictory Beliefs." *Analysis*, 268: pp. 309–319.

Bird, A. (2007). *Nature's Metaphysics: Laws and Properties*, Oxford: Clarendon Press.

Churchland, P., M. (1981). "Eliminative Materialism and the Propositional Attitudes." *The Journal of Philosophy*, 78: pp. 67–90.

Davidson, D. (2001). *Essays on Actions and Events.* Oxford: Clarendon Press.

Davidson, D. (2004). *Problems of Rationality.* Oxford: Clarendon Press.

Deweese-Boyd, Ian, "Self-Deception." *The Stanford Encyclopedia of Philosophy (Spring 2012 Edition)*, Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/spr2012/entries/self-deception/.

Fodor, J., A. (1991). "You Can Fool Some of The People All of The Time, Everything Else Being Equal; Hedged Laws and Psychological Explanations." *Mind*, 397: pp. 19–34.

Friedrich, J. (1993). "Primary Error Detection and Minimization (PEDMIN) Strategies in Social Cognition: A Reinterpretation of Confirmation Bias Phenomena." *Psychological Review*, 2: pp. 298–319.

Johnston, M. (1988). "Self-Deception and the Nature of Mind." *Perspectives on Self-Deception.* In B. McLaughlin and A. O. Rorty (eds.). Berkley: University of California Press: pp. 63–91.

Lazar, A. (1997). "Self-Deception and the Desire to Believe." *Behavioral and Brain Sciences*, 20: pp. 119–120.

Lazar, A. (1999). "Deceiving Oneself or Self-Deceived? On the Formation of Beliefs 'Under the Influence'." *Mind*, 430: pp. 265–290.

Mele, A. (1997). "Real self-deception." *Behavioral and Brain Sciences*, 20: pp. 91–102.

Mele, A. (1998). "Motivated Belief and Agency." *Philosophical Psychology*, 3: pp. 353–369.

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

Mele, A. (2012). "When Are We Self-Deceived?" *Humana Mente*, 20: pp. 1–15.

Mumford, S. (1998). *Dispositions*. Oxford: Oxford University Press.

Pedrini, P. (2010). "Cognition and Desires: How to Solve the 'Selectivity Problem' for Self-Deception." *Pratiche della cognizione, Proceedings of the AISC2010, Italian Society for Cognitive Sciences*: pp. 131–135.

Reutlinger, Alexander, Schurz, Gerhard and Hüttemann, Andreas, "*Ceteris Paribus* Laws." *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, Edward N. Zalta (ed.), URL = http://plato.stanford.edu/archives/spr2011/entries/ceteris-paribus/.

Samuels, R. and Stich, S. (2004). "Rationality & Psychology." *The Oxford Handbook of Rationality*. In A., R. Mele and P. Rawling (eds.). New York: Oxford University Press: 279–300.

Schurz, G. (2001). "What is 'Normal'? An Evolution-Theoretic Foundation for Normic Laws and Their Relation to Statistical Normality." *Philosophy of Science*, 4: pp. 476–497.

Schurz, G. (2002). "Ceteris Paribus Laws: Classification and Deconstruction." In *Ceteris Paribus Laws*, *Erkenntnis*, 57: pp. 351–372.

Smith, M. (1994). *The Moral Problem.* Oxford: Blackwell Publishing.

Trope, Y., & A. Liberman. (1996). "A Social Hypothesis Testing: Cognitive and Motivational Mechanisms." *Social Psychology: Handbook of Basic Principles.* Edited by E. Higgins and A Kruglanski. New York: Guilford Press.