Self-Deception and Agential Authority. A Constitutivist Account

Carla Bagnoli †
carla.bagnoli@unimore.it

ABSTRACT

This paper takes a constitutivist approach to self-deception, and argues that this phenomenon should be evaluated under several dimensions of rationality. The constitutivist approach has the merit of explaining the selective nature of self-deception as well as its being subject to moral sanction. Self-deception is a pragmatic strategy for maintaining the stability of the self, hence continuous with other rational activities of self-constitution. However, its success is limited, and it costs are high: it protects the agent's self by undermining the authority she has on her mental life. To this extent, self-deception is akin to alienation and estrangement. Its morally disturbing feature is its self-serving partiality. The self-deceptive agent settles on standards of justification that are lower than any rational agent would adopt, and thus loses grip on her agency. To capture the moral dimension of self-deception, I defend a Kantian account of the constraints that bear on self-constitution, and argue that it warrants more discriminating standards of agential autonomy than other contemporary minimalist views of self-government.

1. Introduction

There is empirical evidence that self-deception is a quite pervasive phenomenon, even though some would prefer to believe that it is not. Here is an example. Amy knows that her teenager daughter Bea is visibly too thin, does not eat properly, is always concerned with her weight, and selects obsessively

† University of Modena and Reggio Emilia, Italy.

her food; but Amy does not believe that Bea is anorexic. The evidence is accessible and available to her but does not count as a reason for believing that her daughter is anorexic. In fact, she avoids discussing and investigating Bea's eating habits, and other related matters. What does prevent Amy from acquiring the belief that Bea is anorexic? Is Amy irrational, and in what sense? Does her epistemic state bear moral implications, and if so which ones?

Philosophers have given rather different answers to these questions. As a preliminary step, I take self-deception to be the acquisition and retention of a belief despite overwhelming evidence to the contrary. On some views, self-deception is the case where one holds a false belief *p*, possesses evidence that ~*p*, and has some desire or emotion that favor *p*. Intentionalists take self-deception to result from an intention to deceive (Davidson, 1986; Bermudez 2000). Instead, Motivationalists hold that the deception depends on some interfering motivational state, typically a desire or an emotion (Mele 2001; Funkhouser 2003). The self-deceptive agent discounts evidence which one normally would find sufficient to warrant ~*p*, and yet believes *p* instead because of the interference of some desire, emotion or other motivational state favoring *p*.

Self-deception is regarded a case of irrationality, and in extreme cases a pathology that impedes self-knowledge and it is subject to moral sanction. Indeed, this is partly the reason why it is paradoxical. On the one hand, self-deception is a moral charge, which applies to something one does. On the other hand, it implies lack of the relevant sort of self-knowledge that the moral imputability and the applicability of moral sanctions imply. The moral implications of self-deception are to some important extent similar to (interpersonal cases of) deception. But there are some morally relevant aspects of the phenomenon that are absent in interpersonal deception. Arguably, (the charge of) deception implies intentionality, while (the charge of) self-deception does not. For this very reason, lying to oneself is more threatening than being lied to by others, since in the former case it becomes unclear how to protect oneself from deception. We ordinarily assume self-transparency, even though we know that there are large areas of our mental processes and operations that remain inaccessible. One solution is to treat self-deception as a case where our mind is opaque, as it happens for many mental sub-personal processes and

operations.[1] But the interesting aspect of self-deception is that it concerns beliefs and mental states that are normally accessible. Hence, the selective character posits an obstacle to reducing self-deception to a general case of the opacity of the mind because it appears to exhibit some sort of finality. That is, it concerns a selected cluster of beliefs that whose knowledge the agent has an interest in blocking, even though she may not intend to block it.

The question I would like to address is what kind of irrationality self-deception represents, and what moral consequences it carries for the self-deceptive agent. I will argue that self-deception is not merely a pathological phenomenon, but a defensive strategy that is functional to maintaining the stability of the self. As such, the phenomenon of self-deception can be evaluated under different dimensions of rationality. My starting point is to take self-deception as a practical rather than a theoretical phenomenon. Its philosophical relevance resides in the relation the agent bears to her reasons to believe, rather than in the issue of whether she accurately represents the world and her mind as independent objects. I speculate that self-deception is more similar to alienation than to interpersonal deception in this regard. In both cases, self-opacity undermines agential authority, that is, the authority that the agent claims on her action. In focusing on agential authority, I am driven by the conviction that the relevant source of interference in the production and formation process of self-deceptive beliefs is neither an intention to deceive nor a desire, but a concern with one's self-representation. The constitutivist account I am proposing has the merit of explaining the selective nature of self-deception as well as its being subject to moral judgment and sanction. These are very important features of self-deception that elude traditional accounts of self-deception. The motivational explanation fails to fully capture them, and the intentionalist approach treats them inadequately and generates well-known paradoxes about how the agent holds intentionally contradictory beliefs. The proposal is to adopt a constitutivist account of self-knowledge, where agents are responsible for making up their mind, and they are also responsible for self-deception. This is not because self-deception is analogous to deception in that it is brought about by the intention to deceive. Rather, it is because of the special relation (of authority) the agent bears to her own mind and agency.

[1] For a treatment of self-deception that relinquishes the claim about intentionality and say that it self-deception is operated at subintentional level, see Johnston 1988. White (1988) discusses the case of self-deception as an argument for homuncular theories of identity.

Focus on the special responsibility that agents have for their own beliefs helps us see what is wrong with self-deception and why it can be evaluated morally.

This practical account of self-deception has some important consequences about theories of self-knowledge. It belongs to a broadly constitutivist view that takes self-knowledge to be a practical rather than a theoretical matter. A canonical objection against the constitutivist approach to self-knowledge is that it makes the formation and retention of beliefs arbitrary, insofar as it holds that agents make up their mind. Clearly, self-deception would be impossible to describe in any voluntarist view of self-knowledge on which the agent simply decides at whim what to believe, without being subject to any constraint. Constitutivism avoids this problem of arbitrariness by arguing that the formation and retention of beliefs is indeed constrained, hence there are right and wrong ways of constituting beliefs. But constitutivists differ as to what the relevant constraints and their rationale are.[2] The constitutivist account I defend attempts to separate the issue of stability from the issue of autonomy, which is crucial to genuine agential authority. It points out that stability and autonomy are both issues that can be evaluated rationally, but they call into play different dimensions of rationality. My argument is that while self-deception works as a pragmatic strategy to improve or guarantee the stability of the self, it nonetheless undermines its autonomy, hence its authority over action. The self-deceptive agent can even be more stable than the autonomous agent, but it loses authority on her actions. These are all matters of degrees, of course; and one interesting question concerns the scope of self-deception. I will argue that the selective and circumscribed nature of self-deception is crucial to its success as a pragmatic strategy for maintaining stability of the self, which is of limited sustainability.

Furthermore, this argument has some bearing against those theories of agency that either take stability as a property of autonomous agents or take stability equivalent to autonomy. It reveals that the constitutivist views of agency, which hold that agents make up their mind in action, need to lay down some stricter criteria than stability for authorship on mental life. Such criteria

[2] Notably, the main difference concerns the nature of constraints. According to Kantians, such constraints are moral and necessary; for others, they are contingent and their nature is not moral. See Korsgaard 2008, Velleman 2009.

should include moral constraints about how to relate to self and others, and my suggestion is that they be grounded on respect. [3]

2. Self-Deception, Negligence, and Ignorance

It is tempting to think of self-deception as a peculiar case of deception, as if deceiving oneself is analogous to deceiving others. [4] Is not Amy lying to herself, after all? The analogy with deception helps us distinguish self-deception from mere error. The case of Amy who refuses to believe her daughter Bea to be anorexic, despite all the evidence to the contrary, is different from the case in which Greta fails to realize that her son Phil is a drug addict because she fails to recognize and properly collect the evidence, or because she ignores the symptoms of heroin addiction and thus she does not know what counts as evidence in this particular case. Greta may fail to collect the evidence or to adequately interpret the evidence through no faults of her own. Or, she may be utterly negligent. She may simply not care about the whereabouts of her son, and thus refrain from inquiring about his state of health. Or else she may voluntarily disengage from such investigations not out of negligence, but because she does not think it is right of her to intrude and interfere with her son's life, even when his health and prospects are at stake. In these three scenarios, Greta may be holding false beliefs or lacking beliefs about the state of health of her son, but she is not self-deceptive. Self-deception is more similar to deception than to culpable and not culpable ignorance, or error, in this respect.

The selectivity of self-deception is a very important aspect of it. [5] Self-deception concerns only a very specific set of beliefs. In this case, it is only about the mother's beliefs about anorexia, rather than say all beliefs concerning the general state of health of Bea. It is not uncommon that the self-deceptive agent is attentive to all signs but those that matter for the belief she resists. Amy may be quite perceptive of all other aspects of her daughter health, and worried about seasonal cold, while disregarding only the signs of anorexia. The analogy with interpersonal cases of deception helps us see that the self-

[3] This suggestions shows that I tend to side with Kantian forms of constitutivism, but I will not argue directly for any Kantian claim in this paper.

[4] On the moral dimension of the similarity between deception and self-deception, see Baron 1988.

[5] On the so-called selectivity problem, see Bermúdez 1997, 2000.

deceptive agent is deceptive about some particular beliefs, even though she is largely reliable on all other matters. The scope of self-deception is very specific. A massive, global, and self-consistent delusion such as Don Quixote's imaginary world is not self-deception. It requires no internal struggle, no split of the self, and no effort to reach unity. By contrast, the world is always about to intrude in the self-deceptive's existence, and it threatens disaster. There is always a moment where mothers like Amy have to confront reality. Such moments are experienced rather differently than the acquisition of new shattering information about the world. They are likely to be experience of failures, as well as experiences of having failed others. A scenario where Amy eventually realizes that her daughter is anorexic differs significantly from the scenario where she suddenly learns that her daughter is affected by a life-threatening disease.

This asymmetry is often signaled in moral terms. The coming out of self-deception is typically accompanied, or rather, partially constituted by emotions that are appropriate also in the case of moral failure.6 For instance, it is appropriate for Amy to feel guilty for having disregarded the evidence that pointed to Bea's anorexia. Correspondingly, the self-deceptive agent is the target of moral judgments of condemnation or pity, and she is expected to feel guilty upon realization. Perhaps, the moral judgment addressed to the self-deceptive agent is not as strongly negative as the one addressed to the liar, but it is certainly not positive. It is an open question whether the self-deceptive agent deserves to be blamed, and this partly depends on the conditions for being the appropriate target of moral judgment. But it seems largely agreed that the self-deceptive agent morally differs from the one faultlessly lacking relevant information. If the negligent is culpable, the self-deceptive agent is not completely innocent. When she is excused, it is because she is considered a pathological case, less than a fully morally competent agent.

In the case of Amy, the relevant moral implication is that she failed Bea, that is, she failed to pay attention to and take care of her. These are also failures to value Bea as worthy of attention and care, or as I will show next, as failing to recognize Bea as legitimately claiming attention and care. Other cases of self-

[6] I take the category of feelings of guilt to be rather inclusive, and not linked to intentionality. That is, such feelings are appropriate even when the agent did not intentionally cause any harm or violate any moral claims.

deception may not have direct victims as in this case, but there is always something morally objectionable involved. The question is what that is.

3. Feelings of guilt, blame, and pity: the moral relevance of self-deception

It is hard to pinpoint exactly the moral offense of which the self-deceptive agent is guilty. The case is not clear-cut as deception. The deceptive agent manipulates others in order to pursue her own interests or plans. Unlike the deceiver, it is not obvious that the self-deceptive agent intends to deceive herself to further her own interests or plans. In fact, this sounds paradoxical. What further plans and interests does the self-deceptive agent try to pursue despite herself? And how could she pursue some plans in the ignorance of what she herself knows? These are puzzling questions that arise because the analogy with deception leads to thinking of self-deception in terms of intentions. But the analogy can be taken to highlight aspects of self-deception other than its alleged intentionality.

As I take it, the analogy points out that self-deception is a moral charge, associated to some kind of moral sanction. This association, however, does not imply that self-deception is a thoroughly intentional affair. In fact, moral reproach takes different forms whether it is directed to the deceiver or to the self-deceptive agent. It is morally appropriate to blame people who manipulate others in order to get what they want, while pity is a more appropriate moral attitude to address the self-deceptive agent. The moral grammar of feelings of guilt is compatible with the claim that self-deception is not fully intentionally deceptive; and so is the grammar of pity. Nonetheless, self-deception is a case of moral relevance.

Here is the interesting asymmetry, though. In the case of deception, it is apparent who the victim of the moral crime is. In the case of self-deception, instead, things are not simple. The difficulty does not reside only in the fact that it is not obvious whether the self-deceptive agent is culpable of any moral crime, since it is questionable that she intends to deceive. Rather, the difficulty is that it is not clear how to describe the moral offence of the self-deceptive agent. I will argue that there is something morally objectionable about self-deception, even when we put the issue of intentionality aside. There is some self-serving partiality involved in disregarding evidence selectively, which makes the self-deceptive agent look more like the liar than either the negligent

or the ignorant. The reluctance that motivates self-deception is self-concerned. But what does the self-deceptive agent promote, protect, or express?

4. Self-deception as a Practical Phenomenon: a Normative Account

Traditionally, self-deception is seen as an epistemic and theoretical case of irrationality: the self-deceptive agent holds contradictory beliefs about the world. On this description, the problem with Amy is that she believes that Bea is too thin and does not eat properly *and* she also believes that Bea is not anorexic, where these are two contradictory beliefs. The literature on self-deception abounds with strategies to avoid such paradoxical condition (Rorty 1988a, I-II). Countenance of incoherence can take different forms. For instance, some adopt a strategy of temporal partitioning (Bermúdez, 2000). Others, instead, favor the strategy of psychological division, where the self is partitioned into psychological parts that play the role of the deceiver and deceived respectively (Pears, 1984; Davidson, 1985; Rorty, 1988b).[7]

In contrast to these interpretations, I suggest that we describe the case of Amy more like a case where the agent does not take the available evidence to *count as* reasons. This description points to a different aspect of Amy's activity of belief formation. Amy's problem is not that she holds contradictory beliefs, but that she has reason to believe something that she does not in fact believe. Why? The selective nature of self-deception and the analogy with deception discussed above naturally invite us to find answers by investigating further aims of the agent. What does the self-deceptive agent want? What does she try to obtain by lying to herself? Motivationalists respond to these questions by invoking an interfering desire, and treat self-deception as a case of desire-biased belief (Mele, 2001; Nelkin, 2002; Funkhouser, 2005). It is because Amy does not want to be the case that Bea is anorexic. An interesting and illuminating suggestion is that the interfering desire may be not concerned with some state of affairs (a world in which Bea is anorexic), but a self-focused desire (Funkhouser, 2005).[8] What is interesting about this suggestion is that it connects the bias involved in self-deception to the self.

[7] Among the strategies that for avoiding these paradoxes, there are more moderate views about how to draw the division within the self, such as Pears 1984, 1986, 1991, and Davidson 1982, 1985.

[8] Funkhouser (2005) accepts motivationalism, but he interestingly distinguishes between self-focused and world-focused desires, and defends the former account versus the latter.

To further develop this suggestion, however, we need to abandon the talk of desires. I propose a normative model, where the interfering force is neither desire nor an emotion, but a broad normative concern with the agent's own self-representation. I contend that this concern is not reducible to second-order desires about the selves, but it involves appeal to normative ideals of agency, to which the agent holds herself accountable. In the interesting cases of self-deception, this normative concern plays a role in blocking the normative value and weight of beliefs about the agent's not being up to such standards. That is, the self-deceptive agent defends herself against the charge of not being up to her own standards of agency, by blocking the normative force of reasons that support such a judgment. For instance, self-deceptive Amy bracketed or suspended the normative power of reasons for believing that Bea is anorexic. Amy's resistance to form the belief that Bea is anorexic has certainly to do with her desire that Bea be healthy, and with her emotional discomfort of confronting a world where the child is sick, but it has also some more profound connections to how Amy thinks of herself in relation Bea. She knows that Bea is too thin and shows worrisome eating habits, but these considerations have little normative weight in her overall epistemic system. The point is not that the Amy does not access her most intimate thoughts, or that she misses strong evidence about some states of affairs, and thus forms false beliefs or disbelieves what is true. Rather, the key philosophical point in self-deception is that the self-deceptive agent does not take the evidence available to her as reasons. I want to argue that this is a practical mistake, not a theoretical one.

The (practical) problem of how Amy forms her self-deceptive beliefs does not get resolved by endorsing some coherence-driven strategies. More importantly, it is a problem that only Amy can resolve by engaging in practical reflection. In contrast to theoretical reflection about how the world is, practical reflection is driven by the agent's practical concerns. It does not aim at establishing the truth about the world, even though it is constrained by concerns of accuracy and truthfulness. Its purpose is for the agent to determine what she has reason to believe, and this is something that pertains to the context of deliberation.

Here I am invoking a distinction that stands in the background of constitutivist accounts of self-knowledge. Richard Moran has thus formulated the distinction:

> Roughly, a theoretical question is one that is answered by discovery of the fact about oneself of which one was ignorant, whereas a practical question is answered by a decision, and does not arise from ignorance of some antecedent fact about oneself. (Moran, 1988, p. 141).

Accordingly, what it takes to Amy to realize that Bea is anorexic is not some new information about the state of the world, but a change in her practical attitude toward the evidence she already has about Bea. The change is prompted by practical reflection, which does not aim at accuracy in the representation of the world, but it is itself productive of such representations, and driven by a practical concern about what to believe about the world.

To treat self-deception as a practical rather than theoretical issue is not to discount the fact that it is an epistemic condition. Self-deception raises issues about knowledge of oneself. But to capture its philosophical import we should focus on the special relation that the agent bears to her own states of mind. That relation is of authorship. This is the basic claim of constitutivist accounts of self-knowledge. In such accounts, the agent is responsible for what she believes. Hence, she is also responsible for her self-deception. The agent engaged in self-knowledge does not discover some truths about herself through the course of an introspective theoretical investigation; rather, she engages in deliberative activities that are productive of epistemic states.

The epistemic stories that agents elaborate in deliberation are not epistemic stories about themselves as independent objects of knowledge. Such stories are constitutive of self-knowledge. Claiming authorship for what the agent believes of herself is to take responsibility for herself as an agent. Because of its focus on the responsibility for belief, the constitutivist account seems suitable to make sense of two important aspects of self-deception: its selective nature and its moral status. The self-deceptive agent is entitled to feel guilty because she is responsible for her self-deceptive condition. That she is responsible for her beliefs also explains why self-deception is never global or random, but it concerns some beliefs that bear a particular relevance for the self.

5. Self-Deception as a Pragmatic Strategy

It may seem that by making the agent responsible for belief formation the constitutivist account actually dissolves the very problem of self-deception. If it is up to the agent what to believe, how can one discriminate between genuine cases of self-knowledge and self-deception? This is a special case of a general objection against constitutivist views of self-knowledge, which is based on some misunderstanding. Constitutivism does not claim that the agent simply decides what to believe. The claim is not intended to be causal. It is not that the agent brings self-deception about insofar as she decides what to believe. To this extent, the intentionality of the belief is not the relevant philosophical issue. On the constitutivist view, agents are self-interpreting animals. What to believe is something they determine in the first-person, as part of the activities by which they take responsibility for themselves. While belief formation is a practical matter, there are, indeed, norms that constrain and guide its processes.

According to Richard Moran, for instance, such process should respond to criteria of theoretical transparency. One should make up one's mind about *p* on the basis of reasons related to the truth or falsity of *p*. The criteria of theoretical transparency constrain also the formation of attitudes and emotions of fear and love.[9] It is exactly because such constrains hold that we can rationally assess believes, emotions, and attitudes. When such criteria are violated, then the agent makes up her mind for the sake of reasons that are merely *pragmatic*. It seems plausible to treat self-deception as a case where pragmatic reasons prevail, and the agent comes to form and retain beliefs for reasons that are not constrained by criteria of theoretical transparency.

It may seem that self-deception still counts as a case of theoretical irrationality, under this description. My point here is that self-deception is a complex phenomenon and should be assessed according to different dimension of rationality. It is easy to fill in a story where Amy holds very strong pragmatic reasons to discount the evidence she has that her daughter is

[9] «One answers the question of whether to feel hopeful or ashamed by determining whether something is actually hopeful or shameful. Similarly, a practical question about what I want will often be transparent to an impersonal theoretical question about what is good, desirable or useful. It is essential to the rationality of belief that practical questions about it should be transparent in this way» (Moran, 1988, p. 145).

anorexic. Understandably, this is no welcome news. It is normal for parents to think of their kids as safe and perfect, and to resist evidence about their vulnerability. Moreover, suppose that medical research relates anorexia to some deep emotional instability of the anorexic, due to problematic family relations. Perhaps, Amy does not want to confront the possibility that what she represents as a loving nest is not sufficient for Bea's needs. The belief that Bea is anorexic brings along a judgment about herself as a failing mother. To confront this possibility would seriously undermine Amy's own emotional stability. Perhaps Amy lives in a very traditional household where women feel guilty for having a career, even when they do take care of their family, etc. When we take into account the broad deliberative context where self-deceptive beliefs belong, their irrationality is less apparent. Given the full story, it is rational for Amy to discount the belief that Bea is anorexic, because this belief threatens the image she has of herself, and it would undermine her emotional stability. For Amy's own sake, it is preferable to suspend the normative force of the evidence that leads to that belief. It is a rational strategy of defense. What it is threatened is not some particular interest or value that are dear to the agent, but her own understanding of her self. Unlike the liar, the self-deceptive agent does not try to pursue a specific interest or promote an interest. She seems engaged in a much broader and worrisome enterprise. Still, we can recognize some continuity, which can be captured in terms of instrumental or strategic rationality. Self-deception is instrumental to the stability of her self, and to this extent it is a rational strategy. This means that self-deception is not a totally pathological phenomenon. Indeed, its basic processes are continuous with other rational epistemic strategies that underlie correct processes of belief formation. Self-deception is a phenomenon typical and distinctive of animals that hold ideals and representations of themselves. It is because we are self-reflective animals capable of designing representations of ourselves that we are liable to self-deception.[10] The self-deceptive agent is concerned with the

[10] On this aspect see Darwall 1988. To the extent that self-deception requires self-reflection, I agree with Brown (2004) when she holds that the activity of self-emplotment is partially constitutive of self-knowledge and self-deception. But I strongly disagree with Brown's claim that self-deception is a positive epistemic state, for reasons I offer in the text. I have defended a narrative conception of practical identity in Bagnoli 2007. Cf. Holton for a completely different account that takes mistakes about the self to be a necessary condition for self-deception.

coherence and stability of her emotional and epistemic system as any rational agent would be. This is what she is trying to protect. Is she successful?

Debates about this latter question admit only of positive and negative answers. Mele (2001, p. 50) and Nelkin (2002, p. 394) think she is successful in coming to believe as she desires; Funkhouser (2003) thinks she is not. But if I am right to say that self-deception is continuous with normal rational epistemic strategies, the answer should be addressed in a broader context and admit of qualifications.

Baljinder and Thagard (2003) propose that self-deception results from the emotional coherence of beliefs with subjective goals. I think Baljinder & Taghard are right that the self-deceived agent is concerned with improving the emotional coherence of her overall epistemic and deliberative set (beliefs, emotions, and subjective goals). However, this pragmatic strategy is successful only to the extent that it is limited and circumscribed. As a pragmatic strategy to maintain emotional stability and coherence, self-deception is rather limited, and it is important to notice how. First, its success crucially depends on its selective nature. It works only if it is a circumscribed phenomenon. Secondly, because it needs to be so circumscribed, its advantage cannot but be temporary. Typically, as a pragmatic strategy, self-deception comes to an end. Third, when it comes to an end, the self-deceptive agent realizes that stability based on purely pragmatic reasons is not enough, because it fails to afford agential authority, which is necessary to self-knowledge and autonomous agency.[11]

Self-deception is a failure of authorship, which is a dimension of self-knowledge as well as of autonomous agency. The notion of authorship as the capacity to endorse a thought as one's own and justify it on the basis of reasons. Reasons are considerations that make an act intelligible and justifiable. More importantly in this context, reasons convey the relation of authorship. They express a relation between the agent and the action, such that the action can be imputable to the agent as hers. Justifying an action or a belief on the basis of a reason is thus authorizing it and also claiming authorship on it. Hence, actions and beliefs are expressive of one's agency insofar as they are supported by reasons. It is crucial for us that we act and think on the basis of reasons,

[11] That the success of this strategy crucially depends on the limitation of scope is an interesting aspect that makes self-deception similar to deception. The systematic liar is self-defeating as much as the global self-deceiver. After all, Kant was right.

because this is the way we exercise our agency on the world. The threat to our authorship can be more or less tragic and disruptive, depending on the nature of the claims at stake and their relation to our selves. The significance of self-deception varies correspondingly.

6. The Moral Problem of Self-Deception

Self-reflective agents exert a special kind of authority on their mental life. This kind of authority is fundamentally first-personal and, under this reading, it is conceptually linked to (or even identified with) autonomy. Self-deception, I argued, is a pragmatic or defensive strategy for maintaining the stability of the self. Its success is limited, and its costs are high: it protects the agent's self by undermining the authority she has on her mental life. To this extent, self-deception is more akin to alienation and estrangement, and in this final section I propose that we dwell on this similarity in order to appreciate it morally problematic dimension.

According to constitutivist views of agency, the moral person assumes responsibility for herself by regulating her life by her own best judgment. Moral integrity thus amounts to a form of self-government. Rational agents are responsible for the constitution of such self-government. But there are rather different views about how to conceive of self-government. Contemporary accounts of self-government tend to be rather minimalist in terms of the requirement for full authorship and rational self-government.[12] For instance, in his early work, Harry Frankfurt suggested that the upshot of practical reflection aiming at self-government is a «radical separation of the competing desires, one of which is not merely assigned a relatively less favored position, but extruded entirely as an outlaw» (Frankfurt, 1988, p. 170).[13] The aim of this strategy is not so much to resolve the conflict by annulling one desire as to produce a "well-ordered self" by removing the internal obstacle. Interestingly, this aim is achieved by altering the nature of the conflict: once one of the conflicting desires is disavowed, there would be no internal division. Disavowal is a way of disowning some mental state as external, and thus distancing and dissociating oneself from it. Hence, disavowal is not simply a disclaimer; it's an act of choice determining withdrawal of ownership and authorship. The aim of

[12] I borrow this characterization from O'Neill 2004, pp. 13-26.
[13] See also Frankfurt 1988 (pp. 63, 66-67), 2001 (p. 11), 1988 (p. 172), 1999 (p. 136).

self-deception is analogous to the operations of the self that Frankfurt describes as distinctive of autonomous agency. This means that self-deception is more akin to other normal rational activities of the self. But it also indicates that to make sense of its aberration we need to provide stricter constraints than those Frankfurt adopts.

My (Kantian) proposal is to adopt universal criteria of rational scrutiny of reasons.[14] In forming beliefs, adopting attitudes, and take responsibility for ourselves as agent, we should rely on considerations that could be shared by all other rational agents. It is possible to construct such reasons and take them as authoritative if we take ourselves as members of a community of agents with equal standing, governed by norms of mutual respect and recognition. Recognizably, this is a Kantian requirement of practical rationality.[15] It requires that our judgments and actions be intelligible and justified to all relevant others. Insofar as they have equal standing, others are entitled to ask for reasons and accept the burden of offering reasons to us. This is to say that they stand in a relation of mutual recognition with us. While self-knowledge and self-constitution are fundamentally first-personal, they always implicate a broader context of shared norms. Of course, I will not be able to argue directly for this claim. The purpose of these final remarks is merely to point out that in order to distinguish self-deception from other rational epistemic strategies, we need some basic *moral* criteria. Appeal to universal norms of shared rationality explains what is morally wrong with self-deception. The self-deceptive agent does not critically review the considerations that count in favor of beliefs on the basis of shared norms. In order to protect her stability she relies on less demanding constraints.

The morally disturbing feature of self-deception is its partiality.[16] First, it undermines agential autonomy. Out of fear and concern for herself, Amy settles on standards of justification that are lower than any rational agent would adopt, and thus loses grip on her agency. She thereby trades off her autonomy for a limited security and comfort. But she puts herself in no safer place. As we

[14] I develop this view in Bagnoli 2007a, 2007b. See also O'Neill 1985, 2004.

[15] «The concept of every rational being as one who must regard himself as giving universal law through all the maxims of his will, so as to appraise himself and his actions from this point of view, leads to a very fruitful concept dependent upon it, namely that of the kingdom of ends» (Kant, 1785/1996, p. 83.

[16] For a different characterization of what it is wrong with self-deception, see Darwall 1988, Baron 1988.

saw, her pragmatic strategy requires that she be insulated from the world, and this is not possible in the long run. The self-deceptive agent routinely fails herself. Secondly, she fails others. Her partial concern with her safety makes victims. As a result of Amy's self-deception, Bea's desperate call for help is not heard. It may be objected that this happens only in some special harmful cases of self-deception, and it cannot be generalized. But the point is that the self-deceptive agent is inclined to discount reasons that concern others, when such reasons are threatening for herself. It is against this possibility that moral criteria are put forward. Our ordinary epistemic life abounds with small-scale cases of self-deception, and it may seem excessive moral zealousness to treat them as signs of moral failures. The Kantian requirement is not there for the moral fanatic to express her harsh disapproval, but for the reflective agent to prevent that such apparently innocent cases make casualties.

REFERENCES

Bagnoli, C. (2007a). The Authority of Reflection. *Theoria: An International Journal for Theory, History and Foundations of Science*, *22*,(1), 43–52.

Bagnoli, C. (2007b). *L'autorità della morale*. Milano: Feltrinelli.

Baljinder, S., & Thagard, P.R. (2003). Self-Deception and Emotional Coherence. *Minds and Machines*, *13*(2), 213–231.

Baron, M. (1988). What is Wrong with Self-Deception. In B. McLaughlin and A. Oksenberg Rorty (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 431–449.

Bermúdez, J. (1997). Defending Intentionalist Accounts of Self-Deception. *Behavioral and Brain Sciences*, *20*(1), 107–108.

Bermúdez, J. (2000). Self-Deception, Intentions, and Contradictory Beliefs. *Analysis*, *60*(4), 309–319.

Brown, R. (2004). The Emplotted Self: Self-Deception and Self-Knowledge. *Philosophical Papers*, *32*(3), 279–300.

Darwall, S. (1988). Self-Deception, Autonomy, and Moral Constitution. In B.P. McLaughlin & A.Oksenberg Rorty (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 407–430.

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

Frankfurt, H. (1988). *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Frankfurt, H. (1999). *Necessity, Volition, and Love*. Cambridge: Cambridge University Press

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

Holton, R. (2001). What is the Role of the Self in Self-Deception?. *Proceedings of the Aristotelian Society*, *101*(1), 53–69.

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

Kant, I. (1996/1785). Groundwork of the Metaphysic of Morals. In I. Kant, *Practical Philosophy*. (tr. and ed. by M.J. Gregor). The Cambridge Edition of the Works of Immanuel Kant. Cambridge: Cambridge University Press, 37–108.

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

Moran, R. (1988). Making Up Your Mind: Self-Interpretation and Self-constitution. *Ratio (new series)*, *1*, 135–151.

Moran, R. (2001). *Authority and estrangement: An essay on self-knowledge. Princeton*. New Hersey: Princeton University Press.

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

O'Neill, O. (1985). Consistency in Action. In N. Potter, & M. Timmons (Eds.), *Morality and Universality*. Dordrecht: Reidel, 159–186.

O'Neill, O. (2004). Self-Legislation, Autonomy, and the Form of Law. In H. Nagl-Docekal, & R. Langthaler (Eds.), *Recht, Geschichte, Religion: Die Bedeutung Kants für die Gegenwart*. Sonderband der Deutschen Zeitschrift für Philosophie Berlin: Akademie Verlag, 13–26

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