JOURNAL OF
SOCIAL
PHILOSOPHY

# The Attribution of Responsibility to Self-Deceivers

*Anna Elisabetta Galeotti*

## Introductory Notes

The attribution of moral responsibility to self-deceivers is a problematic issue. All accounts of self-deception (SD) acknowledge that the faulty belief-formation process is not under the direct and conscious control of the agent. And according to a common and traditional view of responsibility, it is questionable whether agents are properly responsible for actions out of their own control.[1] The attribution of responsibility to self-deceivers, however, is crucial if one is interested in looking at the practice of SD in the social and political domain.[2] Suppose, for example, that a government becomes self-deceptively convinced that, country is under an imminent nuclear threat by a terrorist group and, on the basis of this belief, carries out a preventive attack on the country that it presumes harbors the terrorist group, with great harmful consequences.[3] If it were the case that no proper responsibility could be attributed to self-deceivers, then SD would turn out to be irrelevant in political and social analysis. In this case, being self-deceived would be actually conflated with being mistaken. Whether mistakes are motivated or not does not significantly change their consideration in social and political analysis. The issue of responsibility is therefore paramount for the possibility of making use of SD as a political and social category.

How to face this issue depends on two intertwining factors: the first is the view of SD and the second is the conception of responsibility. Concerning the first, I shall adopt a view of SD which is intermediate between the two main approaches present in the debate: the intentional account and the causal account. The intentional account considers SD as the product of the subject's intentional strategy, although brought about in a nontransparent fashion,[4] while the causal account views SD as caused by cognitive biases triggered by the subject's motivational state.[5] I hold instead that even though the product of SD, the deceptive belief, is unintentional, the process of belief formation is basically intentional; more precisely, I view SD as a byproduct of intentional steps of the subject elsewhere directed, after the model of invisible hand explanation. I cannot here properly discuss the invisible hand view of SD, and its advantages over either the intentionalist or the causal-motivationist models. For the present argument on responsibility, the relevant and specific feature of the invisible hand model is the conjunction of the unintentionality of outcome with the intentionality of the

JOURNAL of SOCIAL PHILOSOPHY, Vol. 47 No. 4, Winter 2016, 420–438.

process. The agent does not intend to make herself believe the false belief that P, but the latter is brought about by intentional steps of the agent, albeit undirectly.[6]

Concerning the second factor, responsibility, both intentionalists and causalists refer to a view of responsibility focused on control. Those favoring an intentional, although partly unconscious account of SD, usually have little problem attributing responsibility to the agent for his irrationality, although the nonwholly conscious nature of SD and the emotional pressure under which the agent is located may occasionally call for extenuating circumstances in the assignment of blame. But no intentionalist actually doubts that in the SD process there is an agent to whom responsibility can be properly assigned.

By contrast, there might be a problem with the causal account which would appear to free the agent from moral blame, because it is not clear that any agent is involved in SD, hence that there is anyone to attach responsibility to. This, however, is not the prevalent conclusion of the supporters of the causal account. They try to solve the problem by finding a juncture, before or after the causal triggering of biases, where the agent can be said to have control over her volitional or epistemic states. At that juncture, responsibility can be reinstated. Neither the intentionalist nor the causalist solutions, however, seem satisfactory to me, because the first cannot persuasively explain how an unconscious and unaware agent can be in control, and the second how a passive victim of blind causal forces can nevertheless become a controlling agent at certain junctures.

I shall instead argue for a notion of responsibility dispensing with control as the key condition. In recent studies on responsibility, this path has been extensively explored, and control has been substituted by notions such as judgment sensitivity, character expressiveness, or judgment dependence. Generally speaking, such notions rely on character traits for attributing responsibility for actions and attitudes of which the agent was partly or wholly unaware: if the action can be viewed as expression of the "true self" then responsibility to the agent can properly be assigned. This solution, while it dispenses with the condition of awareness of the agent, raises a different issue affecting my view of the self-deceiver. In line with much psychological study of SD, I do not think that SD is necessarily, or even mainly, the expression of a faulty character, one preferring non–truth oriented desires and false comforting beliefs. Instead I argue that, under special circumstances, everyone may in fact become prey to SD. Circumstances more than character set the self-deceptive process in motion. Hence, I shall argue that unaware agents can be held responsible for SD as long as the latter is *their doing*, and not necessarily the expression of their character. In other words, in my argument, agency and not character expressiveness is the key attribution factor.

In a nutshell, this is thesis that I am going to present here. According to the invisible hand view, SD is the agent's doing, admittedly a confused and unaware agent, but not an automaton or a passive victim. The agent's presence is proved by his or her capability of reason-responsiveness ex post, and it provides the key factor for responsibility attribution. In turn, ex post reason-responsiveness opens

the possibility of SD prevention in the future. As the agent has no control over her SD process, it is not sufficient to learn the lesson in order to avoid falling prey to SD in future instances, yet she may recur to the indirect strategies that moral philosophy has devised to deal with the weakness of the will, namely character-building and precommitment. In sum, my proposal presents clear advantages for the consideration of SD in the social and political domain because it not only admits the attribution of responsibility to self-deceivers, but also opens the possibility of SD prevention *via* indirect strategies.

My argument will proceed as follows. First, I shall consider how the issue of moral responsibility has been addressed in the discussion on self-deception. Second, I shall focus on recent views on moral responsibility dispensing with control as the key attribution factor. Third, I shall argue that such views of responsibility, with some critical revisions, fit my understanding of SD as the by-product of intentional doing of the agent. Finally, I shall discuss the related issue of SD prevention and expand on indirect strategies.

## Moral Responsibility in the Literature on Self-Deception

In the discussion on SD, its traditional understanding as lying to oneself has contributed to its consideration as a moral failure for which responsibility could be ascribed without much ado.[7] Such a conception however is highly questionable and much criticized, for how can someone be at the same time the perpetrator and the victim of a lie? This view clearly harbors two paradoxes: the static paradox of believing both P and ∼P at the same time, and the dynamic paradox of making oneself believe what is known to be false. For the intentionalist account to be true, both paradoxes must be explained away as only apparent given that SD cannot be a direct and self-transparent strategy. But if SD is the result of nontransparent strategies, then the intentional account must explain how the nonconscious subject is nevertheless conceived as having proper control over the process.

Given the unsolved difficulties of the intentional account, my discussion is mainly focused on the now prevalent rival model. The causal account rejects the lying to oneself view altogether and conceives SD simply as a motivated belief contrary to the available evidence, produced by cognitive biases triggered by a motivational state. This deflationary approach, while avoiding the complications of the intentional model, presents other problems, among which is the issue of responsibility. Viewing self-deception as causally produced, the subject cannot be held responsible of a mechanism taking place behind her back, so to speak. The supporters of the causal account then suggests that the agent can, however, take control over the input or the output of the mechanism, either checking her emotions and desires before SD has started, so as to preempt their biasing effect on belief formation,[8] or examining the deceptive belief, after its production, in a critical fashion before its final acceptance.

Alfred Mele, an eminent representative of the causal account, favors the first alternative, and moves the locus of responsibility backward from the distorted belief formation process to the motivational set. Agents do not control their cognitive process once the triggering mechanism has set the biasing process in motion, but before that moment they should be capable of exercising control over their desires and emotions for which any rational and moral agent is reason-responsive and accordingly subject of praise or blame. The self-deceiver can then be held blameworthy for having let her unscrutinized motivational state take the lead in the cognitive process.

Neil Levy, by contrast, denies that moral responsibility can be rightly attributed to self-deceivers if SD is conceived as a mere causal event.[9] In order for an agent to be the appropriate target of responsibility attribution,[10] a control requirement must be satisfied in the first place, namely the degree of actual or counterfactual control the agent has over his or her action. But how can we have control over our SD, if SD is viewed as a causal phenomenon? Levy excludes control from the belief formation process and its antecedent motivational state, but considers instead the possibility of locating control *after* SD has taken place, in the final acceptance of the resulting belief.

Briefly, Levy's reasoning is the following. Let us assume that, quite independently from the process of belief formation, we are under the general duty to scrutinize our beliefs critically, as proposed by William Kingdon Clifford's *Ethics of Belief*.[11] Taking Clifford's proposal seriously would mean that before subscribing to any belief, whatever its source, we should preliminarily examine its justification critically and carefully, both concerning the sustaining evidence and the inferential reasoning. In this way, even if we have arrived at the belief that P via a distorted cognitive process, before its final acceptance an independent critical revision should take place anew as an epistemic duty of any responsible believer. Levy, however, acknowledges that such a duty cannot be generally discharged in normal life conditions. Only under special circumstances may the duty hold, and precisely in cases of beliefs concerning relevant matters for moral consideration and/or in cases of those beliefs about which we have doubts. Levy has an easy game, though, to show that under the deflationary account of SD such as Mele's neither condition applies to the self-deceptive P. The first condition has an obvious identification problem: in principle, most beliefs may have implications which are morally relevant, although few have intrinsic moral relevance. And in any case, the moral relevance of a belief depends on whether it plays a role in moral deliberation; but even if the belief is not used in moral reasoning when it is formed, it may be used in the future, blurring the distinction between relevant and irrelevant beliefs for moral considerations. Concerning the second condition, self-deceivers may not experience any doubt, because, under Mele's account, they have come to hold the false belief as the direct effect of a wish causing biases. In sum, if the agent is a victim of a causal process taking place behind his back, it is unlikely that any room is left either for the critical assessment of the belief as morally relevant, or for doubts

about its sustaining evidence, given that both operations would require an agent capable of detaching himself from his SD.

Thus, Levy concludes that there is no room for responsibility ascription under a purely causal account of SD. However, he acknowledges that there may be different explanations of SD according to which the resulting belief is produced unintentionally, yet as a consequence of "cognitively evasive intentional activity."[12] In this case he admits that one of the two conditions above may be met, hence responsibility for SD can be ascribed, but not under a purely deflationary causal account such as Mele's.

Levy's argument is resisted among supporters of the causal view, and other solutions have been explored, because giving up responsibility for SD is not done lightly; it is one thing is to acknowledge that not all SD is culpable, and another to hold that all SD is excusable. Ian DeWeese-Boyd, for example, rejects Levy's line of reasoning and proposes re-examining the possibility of locating control at the beginning of the process, yet in a slightly different fashion than Mele's. He refers to the fact that, in belief formation, data processing should be guided by selective vigilance concerning the threshold of evidence deemed necessary to believe or disbelieve that P. Such a threshold depends on the focal error (either a false positive or a false negative) which the subject wants especially to avoid. In SD, the biasing process has the effect of fixing the focal error on the basis of a wish, which consequently distorts the threshold of evidence necessary to believe that P.[13] According to DeWeese-Boyd, when motivated agents single out the focal error to avoid they are actually reason-responsive and meet the control condition for responsibility to be imputed. But if the motivational state is the general cause of the inaccuracy in data processing, it is hard to think that the selection of the focal error which manipulates the threshold of evidence is instead immune from its causal influence. If that were the case, it would consequently result that agents intentionally chose to let their wish fix the focal error, hence chose cognitive inaccuracy over vigilance, undermining the causal account as a whole. Mele's account, however, excludes that cognitive inaccuracy is entered intentionally by agents, subjecting instead the production of the self-deceptive belief to subintentional causal mechanisms which can hardly be presented as loci of reason responsiveness.[14]

Another way to face the problem of responsibility within the causalist perspective is advanced by Dana Nelkin; her rather complex proposal comprises two distinct suggestions. In the first place, she argues that her specific account within the causalist-motivationist view is more suitable than others, such as that of Mele, for tracking back the motivating desire where ultimately lies the possibility to impute control, hence responsibility. In this respect, her position does not differ from Mele's in that the locus of reason-responsiveness lies in the motivational set triggering the SD process. Her disagreeement with Mele concerns the account of the motivational set, and more specifically the content of the operative wish.[15] A critical discussion of Mele's and Nelkin's disagreement on this point would take us far away from responsibility; here suffice it to say that

Nelkin contends that her view permits an easier tracking of the wish triggering SD than Mele's. Yet, she basically shares Mele's and DeWeese-Boyd's solution to relocate control at the beginning of the process. In this respect, the same critical comments I have advanced above apply to Nelkin as well. In the second place, Nelkin advances a different line of argument based on recent work on responsibility within which she focuses on the consideration of inattention and impetuosity. A line of study has convincingly argued that behavior stemming from inattention is not exempted from responsibility. Given that both inattention and self-deception are conditions out of direct control, if responsibility can be imputed to the first, by analogy it can as well to the latter. I feel much closer to this view, as will become apparent in the next section, yet Nelkin's is a suggestion rather than a proper argument, which among other things leaves unexplained the presence and role of agency in the causal model.

Apart from this latter suggestion, in sum the search for the (vanishing) moment of the process when reason-responsiveness, hence control, can be imputed is doomed in a deflationary and purely causal account. In general, the deflationary account portrays the agent in the SD process as a passive victim of biases. The solution of the responsibility issue advanced by Mele, and not differently by Nelkin, implies that a moment before entering the machine mode, the agent could be reason-responsive for the wish and the emotions triggering the cognitive distortion. The problem with this implication is twofold: on the one hand, the motivational state triggers cognitive biases insofar as it escapes the agent's control; on the other, there is nothing wrong with the agent's wishes and emotions before their triggering of cognitive biases. It is perfectly legitimate to wish one's partner faithful, one's child well-behaved, and one's health fine, just to mention some typical wishes at the origin of SD. Being anxious and worried when appraising unfavorable evidence concerning those wishes is similarly not irrational. What is irrational is the distortion of the cognitive process; but the distortion, being *caused* by the motivational state, is beyond the agent's control. In other words, the agent can be reason-responsive for his desire, but he cannot be reason-responsive for the fact that a perfectly legitimate desire triggers cognitive biases. Therefore, the agent may be perfectly able to give reasons for his desire, but nevertheless unable to provide reasons for the causal effect of his desire on cognition. The solution presented by DeWeese-Boyd instead implies an agent capable of discriminating accuracy and inaccuracy in data processing, and opting for the latter. This solution is however self-defeating, because if the agent were responsible for choosing inaccuracy over accuracy then SD would fall back into the intentional model, and the whole point of the causal model would vanish.

In sum, the attempt to reinstate control at some juncture of the process so as to allow the ascription of responsibility is highly problematic. Even if, according to Levy's proposal, a less deflationary account like mine, where intentional steps by the agent end up in an unintentional false belief, would meet either condition for the duty of belief scrutiny, his approach to responsibility ascription for SD

does not convince me. The self-deceiver is both confused and under emotional pressure; the requirement of control, at whatever moment of SD, seems to me to conflict with the effective agent's powerlessness over the process as a whole. The agent under SD is generally acknowledged to be far from the ideal situation of rational deliberation where the control condition properly applies. Nevertheless, she is asked to exercise control over the impact of emotions and anxiety on cognition from which in the ideal situation of rational deliberation she is supposed to be free. The idea to reinstate control by means of the duty to scrutinize one's beliefs if applied to SD is similarly implausible, even if Levy's two conditions are in principle met. To be sure, often the self-deceiver lingers in his thought and ruminates about evidence, yet in a biased way of which he is not aware and which he does not control; thinking over one's belief in that situation is quite different from rationally scrutinizing one's belief. Generally speaking, how the opacity typical of the self-deceptive belief-formation can then be dispelled and give way to lucid critical thinking about that very belief needs explaining. But granted that self-deceivers have little or no control over their SD, is control the necessary condition for ascribing responsibility to self-deceivers?

## Giving up the Control Condition

The control view is much debated in recent studies on responsibility, and interesting alternatives are on offer. Obviously, no one doubts that actions and attitudes stemming from rational choice, after proper deliberation, are uncontroversial objects of moral responsibility. Yet, our intuitive judgments, ingrained in social life and practices, extend responsibility far beyond actions and attitudes rationally chosen after a proper deliberation; similarly, it is widely acknowledged that much wrongdoing stems from inattention, carelessness, and impulsiveness rather than from vicious and malevolent intentions deliberately pursued in evil strategies.[16] Such considerations have led many scholars to revise the traditional view of responsibility, questioning the control condition as the central attribution key.[17] If control is no longer the central condition, then also choice (and choice-controlled actions) is unnecessary for responsibility attribution, as long as the action or attitude under scrutiny can be viewed as a manifestation of "the true self," or of "judgment sensitivity," or "character expressiveness," or "judgment dependence," to name some of the options replacing control.[18]

Both the control condition and the choice centrality are poignantly criticized by George Sher, whose alternative approach seems to be particularly apt for my case.[19] Control and choice are crucial in what Sher names the "searchlight view" of responsibility, according to which "an agent's responsibility extends only as far as his awareness of what he is doing." Such a view is captured by "the metaphor of conscience as a kind of searchlight."[20] A major problem with this predominant view of responsibility is, according to Sher, the conflation of two distinct and incompatible perspectives on action, namely the

agent's viewpoint at the time of his action, and the ex post detached position of observers (or later self) of the action. The first is internal and forward-looking, the second external and retrospective; only the second is the appropriate perspective for responsibility attribution. It is usually other people (or a later self), from outside, after the action's performance and after its consequences have affected others, that demand reasons and justifications from the agent; and features and considerations about which the agent was not aware at the time of acting are not exempted from demands for reasons. The adoption of this perspective for responsibility ascription seems to me promising and consistent with the practices of the moral community, and the complex critical argument analytically exposed by Sher against the searchlight view seems similarly convincing.

The difficult part of Sher's reasoning concerns his positive alternative to the searchlight view. He must tailor a conception where the scope of responsibility is enlarged but not to the point that agents end up being held responsible even for a heart attack. Thus, he must link the outside judgment of an action as blameworthy with the appropriate internal conditions, granting its origination as an action of the agent. External and internal conditions for an action to be blameworthy are then specified in two disjunctives:

> The agent "is responsible for his act's wrongness or foolishness if and only if he: 1) is aware that the act is wrong or foolish when he performs it, or else 2) is unaware that the act is wrong or foolish *despite having evidence* for its wrongness or foolishness *his failure* to recognize which a) *falls below some applicable standard*, and b) is caused by the interaction of some combination of his *constitutive attitudes, dispositions and traits*."[21]

Sher's definition captures the idea that in the case of unawareness, responsibility for wrongdoing or foolish acts cannot be withheld if (i) the agent had the evidence necessary for understanding the foolishness or wrongness of his act at the time of acting and (ii) nevertheless failed to grasp it; (iii) thus falling below standards. (i–iii) jointly represent the outside perspective of someone asking the agent ex post: "How could you forget our meeting?" If the agent can answer this question saying, "I was kidnapped and imprisoned in a basement for three hours," then his apparent misdeed is definitely excused because his failure to turn up at the meeting was not an *omission by the agent*, but an *event* caused by external factors. No excuses are however acceptable if (iv) the agent's failure is dependent on constitutive traits, attitudes and dispositions of his. This last condition, as specified in (2b), grants the origination problem of the act: it grants that the act is performed by the agent and not caused by some contingent or extrinsic factor, such as coercion, brainwashing, manipulation, or reflexive behavior. To put it differently: whenever the agent fails to grasp the evidence for seeing his action as wrong or foolish and the failure can be attributed to him and not to some external or contingent impairments, then he can and ought to be held responsible for his action.

According to this description, however, it is not obvious that agents can be held responsible for their SD. SD is not generally caused by a character vice, but is instead the quite common product of a motivated form of belief formation that strikes any normal cognizer under certain circumstances. Experimental psychology has amply proved that this is the case.[22] This fact does not exclude that certain character dispositions may favor certain agents to fall more easily or repeatedly prey to SD, yet it excludes the general or predominant dependency of SD from character traits. But in this case it is not clear that disjunctive (2b) is met in most SD cases.

The fact is that, although SD is not predominantly or specifically produced by constitutive traits of the agent's character, nonetheless there is no doubt that it originates in agents, in particular situations and circumstances. It is not that the circumstances directly cause SD, bypassing agency, but that attributing SD responsibility to character expressiveness misconstrues how the faulty belief is formed. Under the invisible hand account, the subject is not just a passive automaton, nor the victim of causal mechanisms, yet she is in a condition of both opacity and epistemic confusion: SD is the typical instantiation of faulty mental activities whose faultiness escapes the subject's awareness, despite available evidence to the contrary. It actually fits the (i–iii) features above. In other words, it would seem a typical instantiation of those acts performed "in the dark" and yet proper objects of responsibility assignment for which Sher has proposed his revised view. Not being traceable to character expressiveness, hence not fitting the (iv) feature, it seems that either responsibility is not attributable to self-deceivers or that Sher's characterological condition must fail.

I favor the latter option, and not only because I am reluctant to lift responsibility from the self-deceiver's shoulders. The reason Sher has posited character expressiveness lies in the need to link the outside perspective with the internal one; that is, to solve the origination issue. But the fact that something originates in me, instead of happening to me, does not require that my *constitutive character* (*or true self*) be the originator, especially not in cases of foolish or careless actions. The latter are not necessarily the product of flawed characters; given the intrinsic vulnerability of human rationality and agency, all of us occasionally act absentmindedly, foolishly, unthinkingly, especially if pressed by time, anxiety, or other emotions impairing cognitive lucidity. All of us forget something now and then, and although some people may be more forgetful than others, that fact does not alter the ascription of responsibility for any act of inattention. If John forgets the train ticket for whatever reason (usual absentmindedness, haste, or an untimely call while he was preparing his bag), he knows that he cannot be allowed to travel without a ticket simply because he did not do it on purpose. But John's acknowledgment of his foolishness and the consequent acceptance of responsibility and blame do not depend on his forgetfulness being expressive of his character. Whether or not it is his habit to be absent-minded, he is in any case responsible because he and nobody else omitted to put the ticket in the bag, and the omission was properly his. I am not claiming that forgetfulness is never

a sign of a flawed character; I am claiming that *responsibility for one's forgetfulness* does not derive from being the expression of a flawed character.

Philosophers analyzing responsibility are obviously concerned that it is the agent—and not an external force, a blind mechanism, or an evil genie—who produces the outcome, albeit inadvertently or absentmindedly, and if the subject is the agent and not the victim of the action, then responsive attitudes of blame properly apply. Yet, in order to assign blame for a misdeed or a misjudgment to an agent, it is sufficient that there is an agent, and not a victim or an automaton. For a misdeed to be a misdeed of the agent it need not express her or his character. The inadvertent wrongdoer may not be at the top of his or her agential authority, yet he or she is an agent for all that; human imperfection of rationality and morality pertains to human agency, and the lapses of either are the agent's, expressing general human vulnerability more than character imperfections.

In sum, the originating condition requires agency as opposed to mechanism, automatons, and passive victims, but it does not require a reference to the true self. I would stick to a conception that ascribes the origination issue to the simple presence of agency, a presence that in the absence of coercion, brainwashing, manipulation, external circumstances, or physical or mental disabilities must be presumed. In other words, I adopt a conception of responsibility where ascription does not depend on the control condition ex ante, but rather by the fact that an action, although performed without full awareness, is performed by an agent capable of responding to the question, "how could you do it?" Responsibility is ascribed because of the presence of agency in the act, a presence which is proved by the reason-responsiveness of the agent ex post. Only a moral agent can produce a self-judgment ex post, acknowledging past conduct as *his* and *faulty*. Actually, it is not the case that reason-responsiveness can retrospectively reinstate the agent's control on his past conduct. This would be the case if the control condition were conceived of as counterfactual. But I think that Sher is right in saying that control cannot be counterfactually imputed, for failure of control is a non-event unrelated to the subject.[23] It seems indisputable that, at the moment of acting, the subject's agential capabilities were substandard, so to speak, but if ex post the agent were not reason-responsive, this fact would suggest either that agency was absent in that instance or that the subject has a more serious agential deficit. Ex post reason-responsiveness is required to guarantee that the person has agential capability in general, and that her conduct or act has been produced by a local failure of her agential capabilities to be ascribed to herself and not to external factors. In this case, SD is comprised in the range of (mental) acts for which responsibility is properly ascribed.

In turn, ex post reason-responsiveness is what allows the agent to learn the lesson, and to be more careful, prudent, and considerate in future similar occurrences. Learning a lesson is not the *reason why* we ascribe responsibility for past actions, but it is the *sign* that a certain conduct was performed by an agent, who is able to reflect on her deeds or misdeeds ex post, and possibly avoid similar mistakes in the future. Suppose that, instead, the objectionable behavior turns

out ex post as nonimputable to the agent because it was caused by illness or by external impediments. In such a case, there is no lesson to learn for the agent by an ex post reflection on behavior. If Anna did not show up at the meeting because she fainted or because she was stuck in the elevator, there is little she can learn by a retrospective reflection in order to avoid similar occurrences in the future. Responsibility can be attributed to out-of-control acts, yet in such cases, given the practice of exchanging reasons among moral agents, if the act was the agent's and responsibility can be attributed, then she is expected to learn the lesson for the future. It is not by chance that relapses of misbehavior are usually regarded as more blameworthy than a one-time misdeed. If there is no expectation about learning the lesson, that is because the behavior is not considered as performed by the agent, but as caused by some external factors. In this respect, the possibility of learning a lesson ex post is the signal of agency. In other words, we can judge conduct as blameworthy or foolish if performed by *agents* for which *control* can be learned. If there is nothing to learn for the subject, I doubt that responsibility can be assigned. In sum, the internal condition is met if the action for which responsibility is to be attributed is an action of the agent, and the agency condition is in turn specified, negatively, by the absence of coercion, brainwashing, manipulation, and external causes, and positively by the fact that the agent is ex post reason-responsive, hence can learn a lesson and acquire control of future conduct.

How does SD fit with this view of responsibility attribution? As noted above, responsibility can be attributed when someone performs a wrong or foolish act not consciously in the following conditions: (a) the agent has evidence that the act is wrong or foolish at the time; (b) she fails to recognize that evidence; (c) the failure falls below some applicable standards; (d) the faulty act is imputable to the agent. In turn, (d) implies that no completely incapacitating conditions are present and that the agent is ex post reason-responsive; that is, she can understand her fault and learn the lesson. These conditions perfectly apply to the invisible hand account of SD; the agent is unaware of the faultiness of what she is doing, failing to consider the available evidence properly. As a result, her mental activities are below standard, and yet are performed by an agent, albeit opaque and confused.[24] The agent is under emotional pressure, to be sure, but such circumstance is not incapacitating to the point of erasing agency; and ex post, exiting SD, the agent usually recognizes the faultiness of her condition and feels regret and shame at having fooled herself. The ex post reason-responsiveness of the previous self-deceiver is present. Hence, the self-deceiver can properly be held responsible for her deceptive belief.[25]

It is, however, less clear that reason-responsiveness ex post can help the agent to avoid future episodes of SD. Agency involved in SD exhibits the structural limitations and imperfections of human thought with limited control on mental activities, and in the SD process the agent is not aware of becoming self-deceived. When the favorable conditions for SD obtain, in principle the subject may recognize them and resist the impact of emotions and desires, preventing

the process to start. Such a hypothetical option however is hardly available in practice, because the possibility of avoiding future SD even in a disenchanted subject meets many obstacles on its way. Desires and emotions are resisted by internal argument, but such argument is in turn exposed to the motivational influence and to biases, at different junctures. Moreover, when we move from the realm of personal life and interpersonal morality, the blame for SD does not simply concern its foolishness and impact on the agent's self-respect. There is at stake not simply the lapse in one's agential authority and autonomy, but much heavier moral considerations having to do with harming and wronging other people. In that case, we cannot rest content with assigning stronger moral condemnation and culpability: we should seek preventive measures.

## SD Prevention

If we move from the realm of moral theory to the social and political domain, then the consideration of SD's harmful consequences on other people is crucial. If momentous decision-making is based on self-deceptive beliefs, as I contend is sometimes the case in foreign policy, then the resulting policy is grounded on false premises, hence usually flawed, and the harmful consequences are far-reaching.[26] The issue of responsibility ascription must thus expand to that of SD prevention. With reference to prevention, SD in fact presents an advantage over the two phenomena with which it is often equated: straight lying, on the one hand, and honest mistakes, on the other. Deception and mistakes can neither be predicted nor prevented. By contrast SD is in principle open to prevention. Contrary to either lying or mistakes, SD is triggered in specific circumstances, namely when certain motivations meet with contrary evidence and the costs of inaccuracy are low or can be ignored. Prevention depends, first of all, on the possibility of detecting such favorable circumstances for SD to occur. It may be that under these circumstances people with certain characters fall prey to SD more easily, but from the viewpoint of social and political analysis the crucial feature to predict SD and prevent its harmful consequences is constituted by these favorable circumstances. Consider, moreover, that in the political domain most cases of SD worthy of analysis are actually collective products, as shown by the well-known studies on groupthink,[27] and hence character considerations become irrelevant and circumstances paramount.

The second element to consider for some prevention strategy is that the agent is not in control of SD: she is not openly and willingly performing her deception and her thought is fogged by motivations and liable to biases. Even after acknowledging the foolishness and the potential harm of SD, the agent may yet be unable to access the reasons against SD when the pertinent circumstances obtain, and emotional pressure may darken awareness yet again. I would say that direct control over one's SD, although not impossible in principle, is difficult and rare in practice. But direct control is not the only form of control

which we have at our disposal in regulating our actions and beliefs. Moral psychology has singled out at least two forms of indirect control precisely in order to bypass potential weakness of the will: character-building and precommitment.

Character-building is the Aristotelian strategy of fortifying one's conduct with virtues which provide the dispositions to act morally. Virtues follow from a right character which, in turn, is molded out of good habits and discipline. The latter are necessary to prevent agents from falling prey to weakness of the will.[28] Precisely from Aristotle, George Ainslie picked up the idea of discipline to win the internal conflict between short-term and long-term interests, where morality can be seen as a long-term project imposing itself on immediate rewards.[29] And like Aristotle, Ainslie believes that only learning to be disciplined by the appropriate habits can secure the agent's commitment to long term-interests and morality. Good habits and discipline comprise the proper focus of moral teaching and learning. We have already seen that ex post SD can engender moral learning in agents, yet the character-building project is of limited efficacy as far as SD is concerned. Ainslie's account helps us to understand why: according to his view, self-discipline is acquired to gain immediate rewards from our long-term projects. With reference to knowledge, the default motivation is to believe what we wish according to our desires and fantasies, and the acquired motivation, regulated by self-discipline, is truth-seeking and knowledge constrained by evidence. Lapses of self-discipline, however, are to be expected under certain circumstances, namely abundance of rewards and/or crumbling of future plans. If his account is accurate, it is not the case that the self-deceiver is someone who lacks self-discipline and character; if that were the case, the subject would be delusional instead of self-deceived. The self-deceiver is rather someone experiencing a lapse of self-discipline under special circumstances, lapses that must be actually expected by the beings that we are.[30] Certainly, agents can improve their self-discipline and fortify their prudential motivations, and yet character-building alone cannot be depended upon to prevent SD, especially when the focus is on the social and political domain, and when most relevant episodes of SD are collective.

Alternatively, for more effective prevention of SD, precommitment is in order. Precommitment is the strategy symbolized by the story of Ulysses and the sirens. In order to be able to listen to the sirens' song, without jumping off board and drowning in the sea, the clever Ulysses fills his crew's ears with wax and orders them to tie him to the ship's mast. When the temptation comes with the spellbinding sound, his sailors are safe with their hearing blocked and he is prevented from following the urge to jump off the ship by being tied up. What Ulysses did was to create a constraint on options at time $t^1$, under conditions of cognitive lucidity, so as to avoid, at time $t^2$, under emotional pressure, falling prey to a temptation that one knows is difficult to resist.[31] Precommitment is thus the rational strategy to control one's lapses of discipline and rationality in difficult situations. Yet philosophers regard this solution as less than desirable,

because an admission of defeat and weakness for the relevant behavior at time $t^2$ would not be autonomously chosen, nor follow from rational choice, but would be forced by external constraints which bind individual freedom.[32] Even if at time $t^1$ constraints have been autonomously chosen by the agent, at time $t^2$ he is bound, so that his behavior is not *his action* properly. Along this line of reasoning, precommitment is redescribed as an intentional manipulation of one's autonomy for bringing about a desired outcome, but also as a decrease of overall autonomy. A proper discussion on precommitment as a general rational and moral strategy is however beside the point here. The point is rather the actual and feasible prevention of SD. If SD is a bad and morally objectionable state for the self and potentially dangerous for others, and is a proper object of blame although not under the agent's direct control, what feasible remedy can be devised, if character resoluteness is empirically insufficient to prevent SD? Precommitment looks like a suitable candidate, although it may be defective in terms of ideal rationality and morality. Yet I think that the alleged moral defectiveness of precommitment does not apply to our case, for SD is an autonomy-impairing state, and avoiding SD through self-binding neither decrease the subject's overall autonomy nor his liberty, as will become apparent in examining how it may work.

We have seen that the agent hardly has control over the process of bringing about SD, not only because it is difficult to resist temptation under emotional pressure, but also and especially because the agent is unaware of what is going on in her beliefs. And yet afterwards, the agent has reason to regret her previous deceptive state, hence has reason to try to avoid the same mistake in the future. She has learned that when confronting negative threatening evidence, potentially disruptive for the self, and about which she feels basically powerless, the shortcut of SD is a ready option. How can one avoid falling prey to SD then? How can one bind herself to truth-seeking as Ulysses made himself bound to the ship mast?

In our case, precommitment may work by trusting oneself to a "referee" concerning one's motivated hypothesis. I am here suggesting a reversal of what usually happens in SD cases. As widely acknowledged, SD is often supported and kept going by a charitable community.[33] More often than not, the self-deceiver tells some friends about the deceptive hypothesis so as to be reassured and have it implicitly confirmed and usually friends detect SD, but also see it as a "vital lie" for the agent and do not feel like awakening her abruptly, especially because they see her as powerless.[34] They do not think that their implicit assent will constitute an important piece of evidence for the self-deceiver in favor of her false belief, and will reinforce her deception. In the precommitment strategy that I am suggesting here, the collusive community should instead be transformed into a refereeing one to help guide the subject away from the SD shortcut. It should, however, be the agent herself who directly and explicitly confers on her special friend(s) the authority of referee(s) over her potential SD. Such authorization is important because, in the first place, the friends of the

prospective self-deceiver should avoid the self-appointed role of guardians, with its implicit self-righteousness and paternalism, and in the second place, the agent ought to take responsibility for their intervention in order to fully subscribe to her (pre)commitment against SD. Moreover, only someone with the explicit authority to speak her mind clearly and truthfully can expect to be taken seriously by the authorizing agent, despite the difficult circumstances and the emotional pressure in favor of the soothing deceptive belief. Conversely, the agent can take credit for SD prevention only with an explicit authorizing agreement, made ex ante, under conditions of cognitive lucidity. Once such agreement is in place, the agent becomes fully responsible, hence more heavily blameworthy if she dismisses the referee's advice, while if she takes the advice she will have the full credit for SD avoidance. The remedy for SD is reached by a detour through the assistance of an authorized referee to bind the agent to a clear-sighted interpretation of the evidence. This binding neither diminishes the subject's liberty nor decreases her autonomy because, on the one hand, avoiding SD means enhancing her autonomy as rational cognizer, and on the other hand, after all, it is still up to the agent to follow the friends' advice at time $t^2$.

The strategy of precommitment can moreover be used at the institutional level as a prophylactic measure against the SD of presidents, cabinets, and crucial decision makers. If we admit that sometimes crucial decisions and policies are based on deceptive beliefs influenced by motivations instead of epistemic principles, the issue of SD prevention acquires a special urgency.[35] How to design an acceptable institutional form of precommitment is complex and should be dealt with by constitutional experts. Yet in principle we can imagine that independent observers (perhaps a parliamentary committee) may act as overseers of governmental decisions when the decision is momentous, and especially when information is classified and hence more easily open to bias induced by wishes and preferred courses of action. In these circumstances, an external and unbiased consideration of data may provide that devil's advocate that so many participants have acknowledged to be crucial to avoid loops of faulty reasoning. I am well aware of the many pragmatic difficulties of translating this idea into a viable institutional option, but difficulties are not a reason to dismiss the idea of SD preventive measures and precommitment.

## Conclusion

I have explored the moral significance and implications of SD specifically focusing on the issue of responsibility. In this respect, there seems to be a major problem specifically with unintentional accounts: can someone in the grip of emotionally loaded desires be held responsible for the biasing of her thoughts, which took place causally and behind her back? This question has led me first to examine how scholars of SD have faced this problem. Following the traditional view of responsibility informed by the control condition, they have mostly attempted to single out a juncture in the process of the self-deceptive belief

formation where control may plausibly be imputed and hence responsibility attributed. I have argued that these solutions are unsatisfactory, because in addition to being highly implausible in point of fact they are conceptually muddled, given that the agent is portrayed as at the same time a passive victim of biases and a resolute self-knower. Consequently, my analysis has turned to different views of responsibility dispensing with the control condition.

By a critical examination of Sher's theory, I have provided an account of how self-deceivers can be held responsible for their act, even if they did not perform it intentionally and knowingly. From my perspective, the attribution of responsibility is linked to the possibility of future prevention, by coming to see SD as bad and wrong. Having reasons recommending the avoidance of SD, although, may not be sufficient for its future prevention. Within the moral tradition, we can find some suggestions for dealing with desirable states which cannot be reached directly. A first suggestion comes from Aristotle, who sees the solution to *akrasia* in the process of character-building, by means of good education and healthy habits. Yet it is not clear whether this strategy can effectively work against SD, especially in the social and political domain. The second suggestion, precommitment, seems more promising. This strategy requires some moral learning as well: the agent, reflecting on her previous SD, should come to the conclusion that SD is detrimental for her and others, all things considered, and hence should be avoided. But then, in order to carry out her resolution, the agent should invest some friend(s) with the authority of referee(s) when typical circumstances for SD arise. The referee(s) should represent the external reasonable point of view concerning the negative evidence, and highlight the motivated biases in the agent's belief. Obviously, such precommitment against SD does not represent a physical constraint like it was for Ulysses who was literally bound. The agent can reject the referee's advice. Nevertheless, if the referee has been explicitly authorized, the natural tendency to defend one's unwarranted belief will be reduced by the very act of authorization freely subscribed to by the agent, under conditions of cognitive lucidity. Moreover, the very same authorization also allows the agent to take credit for the avoidance of SD. And, if the strategy is successful, one can imagine that good habits, concerning evidence processing, will be reinforced and character fortified.

*This article has been presented in the Public Reason Seminar of my Department and then at the University of Modena. I thank Carla Bagnoli, Ian Carter, and Valeria Ottonelli for their comments. I also thank two anonymous referees for their remarks.*

## Notes

[1] See Fischer and Ravizza (1998), Levy (2004), Sher (2006).

[2] I have argued in favor of making use of SD in politics and more specifically in certain momentous decision making processes (Galeotti 2015).

[3] The example is not simply a case of philosophical fiction given that Stephen Holmes (2007) argues that it is precisely what happened in the aftermath of the 9/11 terrorist attack: both the government and the American people, as an effect of a self-deceptive process, became convinced of being under an imminent nuclear threat, and such a conviction was then the grounds for the intervention in Afghanistan first and in Iraq later.

[4] As examples of the intentional approach, see Pears (1984), Davidson (1985), Talbott (1995), Fingarette (1998), Bermúdez (2000).

[5] As examples of the causal approach, see Mele (1997, 2001), Lazar (1997), Scott-Kakures (1996).

[6] A proper presentation of the model is discussed in Galeotti (2012).

[7] See Demos (1960), Foss (1980).

[8] Mele (2001, 103), Barnes (1997, 83).

[9] Levy (2004).

[10] Fischer and Ravizza (1998).

[11] Clifford (1886).

[12] Levy (2004, 299).

[13] DeWeese-Boyd (2007).

[14] Mele (2001).

[15] Nelkin has presented her version of the causalist-motivationist account in Nelkin (2002). Her account is based on the definition of the wish setting in motion SD as the "desire to believe that P" instead of "the desire that P," and on stressing that such desire makes her account "content-restricted" concerning the operative wish in SD.

[16] Jenni (2003), O'Hagan (2012).

[17] Sher (2006).

[18] For alternatives to the control condition, see Wolf (1990), Scanlon (1998), Sher (2006), Smith (2008).

[19] Sher (2009).

[20] Sher (2009, 6).

[21] Sher (2009, 88).

[22] Sackeim and Gur (1985), Wentura and Greve (2003).

[23] Sher (2009, 85).

[24] This is where the invisible hand model of SD presents a clear advantage over the causalist account given that, according to the latter, SD is a causal event that befalls the agent. The causalist must then account how, despite being victims, self-deceivers are nevertheless also agents, and not in general, but precisely in the process of SD formation. By contrast, in the invisible hand model SD is a doing of the agent ending up in an unintended outcome.

[25] My suggestion to link agency on SD with reason-responsiveness ex post resonates with an argument made by John Christman (2009, 142) concerning autonomy. He states that the suspension of reason-responsiveness of the agent in certain states (for instance when being madly in love) is not sufficient to consider that person as nonautonomous, as long as the state in question is not produced by brainwashing, manipulation, or hypnosis.

[26] In Galeotti (2015), I have discussed some such cases, for example the Bay of Pigs failed invasion.

[27] Janis (1982).

[28] Aristotle ([350BCE] 2009).

[29] Ainslie (2001).

[30] Following Ainslie's view, the circumstance for SD would specifically be the negative evidence concerning a crucial wish of the subject which prospectively makes long-term projects meaningless. The crumbling down of future plans induces a lapse in the self-discipline concerning knowledge and the recourse to self-rewards, namely deceptive and comforting beliefs.

[31] Elster (1979; 2000).

[32] Not all views of autonomy share such a judgment. Christman's view, for example, rescues precommitment from the accusation of being a nonautonomous strategy (Christman 2009).

[33] Rorty (1996).

[34] On the "vital lie" view of SD, see Rorty (1988), Vaillant (1993), Goleman (1985), Lockard and Paulhus (1988).

[35] On this point, there is relatively little study, but I shall refer to Daniel Goleman's interpretation of the Bay of Pigs (Goleman 1985) and to Holmes's interpretation of the widespread belief of nuclear threat after 9/11 in the U.S. government and the people (Holmes 2007) both as episodes of SD.

## References


Ainslie, George. 2001. *Breakdown of Will*. Cambridge: Cambridge University Press.

Aristotle [350 BCE] 2009. *Nicomachean Ethics*, trans. D.W. Ross. Oxford: Oxford University Press.

Barnes, Annette. 1997. *Seeing through Self-Deception*. Cambridge: Cambridge University Press.

Bermúdez, José Luis. 2000. "Self-Deception, Intention and Contradictory Beliefs." *Analysis* 60: 309–19.

Clifford, William Kingdon. 1886. *"The Ethics of Belief." In Lectures and Essays*, ed. Leslie Stephen and Frederick Pollock, 339–63. London: Macmillan.

Christman, John. 2009. *The Politics of Persons: Individual Authority and Socio-historical Selves*. Cambridge: Cambridge University Press.

Davidson, Donald. 1985. "Deception and Division," In *Actions and Events: Perspectives on the Philosophy of Donald Davidson*, ed. Ernest LePore, E. and Brian McLaughlin, 138–48. Oxford: Basil Blackwell.

Demos, Raphael. 1960. "Lying to Oneself." *Journal of Philosophy* 57: 588–95.

DeWeese-Boyd, Ian. 2007. "Taking Care: Self-Deception, Culpability and Control." *Theorema* 26: 161–76.

Elster, Jon. 1979. *Ulysses and the Sirens*. Cambridge: Cambridge University Press.

——. 2000. *Ulysses Unbound*. Cambridge: Cambridge University Press.

Fischer, John Martin, and Ravizza, Mark. (1998) *Responsibility and Control. A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Fingarette, Herbert. 1998. "Self-Deception Needs No Explaining." *Philosophical Quarterly* 48: 289–301.

Foss, Jeffrey E. 1980. "Rethinking Self-Deception." *American Philosophical Quarterly* 17: 237–43.

Galeotti, Anna Elisabetta. 2012. "Self-Deception: Intentional Plan or Mental Event?" *Humana Mente* 20: 41–66.

Galeotti, A.E. 2015. "Liars or Self-Deceived? Reflections on Political Deception." *Political Studies* 63: 887–902.

Goleman, Daniel. 1985. *Vital Lies, Simple Truths: The Psychology of Self-Deception*. New York: Simon & Schuster.

Holmes, Stephen. 2007. *The Matador's Cape: America's Reckless Response to Terror*. New York: Cambridge University Press.

Janis, Irving, L. 1982. *Groupthink: Psychological Studies of Policy Decisions and Fiascoes*. Boston: Wadsworth.

Jenni, Kathie. 2003. "Vices of Inattention." *Journal of Applied Philosophy* 20: 279–95.

Lazar, Ariela. 1997. "Self-Deception and the Desire to Believe." *Behavioral and Brain Sciences* 20: 119–20.

Levy, Neil. 2004. "Self-Deception and Moral Responsibility." *Ratio* 27: 294–311.

Lockard, Joan S., and Paulhus, Delroy L., eds. 1988. *Self-Deception: An Adaptive Mechanism*. Englewood Cliffs. NJ: Prentice-Hall.

Mele, Alfred R. 1997. "Real Self-Deception." *Behavioral and Brain Sciences* 20: 91–102.

——. 2001. *Self-Deception Unmasked*. Princeton, NJ: Princeton University Press.

Nelkin, Dana K. 2002. "Self-Deception, Motivation and the Desire to Believe." *Pacific Philosophical Quarterly* 83: 84–406.

O'Hagan, Emer. 2012. "Self-Knowledge and Moral Stupidity." *Ratio* 25: 291–306.

Pears, David. 1984. *Motivated Irrationality*. Oxford: Oxford University Press.
Rorty, Amélie Oksenberg. 1988. "The Deceptive Self: Liars, Layers and Lairs." In *Perspectives on Self-Deception*, ed. Brian P. McLaughlin and Amélie Oksenberg Rorty, 11–28. Berkeley: University of California Press.
——. 1996. "User-Friendly Self-Deception." In *Self and Deception: A Cross-Cultural Philosophical Enquiry*, ed. Roger T. Ames and Wimal Dissanayake, 73–89, Albany: State University of New York Press.
Sackeim, Harold A., and Gur, Ruben. 1985. "Voice Recognition and the Ontological Status of Self-Deception." *Journal of Personality and Social Psychology* 48: 1365–68.
Scanlon, T. 1998. *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
Scott-Kakures, Dion. 1996. "Self-Deception and Internal Irrationality." *Philosophy and Phenomenological Research* 56: 31−56.
Sher, George. 2006. "Out of Control." *Ethics* 116: 285–301.
——. 2009. *Who Knew? Responsibility without Awareness*. Oxford: Oxford University Press.
Smith, Angela M. 2008. "Control, Responsibility and Moral Assessment." *Philosophical Studies* 138: 367–92.
Talbott, W. J. 1995. "Intentional Self-Deception in a Single, Coherent Self." *Philosophy and Phenomenological Research* 55: 27–74.
Vaillant, George E. 1993. *The Wisdom of the Ego*. Cambridge, MA: Harvard University Press.
Wentura, Dirk., and Greve, Werner. 2003. "Who Wants to Be Erudite? Everyone! Evidence for Automatic Adaptations of Trait Definition." *Social Cognition* 22: 30–53.
Wolf, Susan R. 1990. *Freedom Within Reason*. Oxford: Oxford University Press.