Can You Succeed in Intentionally Deceiving Yourself?[*]

Dion Scott-Kakures[†]
dion.scott-kakures@scrippscollege.edu

ABSTRACT

According to intentionalists, self-deceivers exercise the sort of control over their belief-forming processes that, in standard cases of interpersonal deception, the deceiver exercises over the deceived's belief forming processes – they intentionally deceive themselves. I'll argue here that interpersonal deception is not an available model for the sort of putatively distinctive control the self-deceiver exercises over her belief-forming processes and beliefs. I concentrate attention on a kind of case in which an agent allegedly intentionally causes herself to come to have a false belief. I hope to show that contrary to appearances, the agents in such cases do not intentionally cause themselves to have false beliefs – do not intentionally deceive themselves. Indeed, if we take the model of interpersonal intentional deception seriously, we ought to conclude that a self-deceiver, so regarded, deceives herself *unintentionally*.

1. Introduction

We are all familiar with the unhappy fact that we frequently deceive ourselves – cause ourselves to have false beliefs. If this sounds hyperbolic or alarming, it should be recalled that, typically, we cause ourselves to have false beliefs in unintentional fashion.[1] Aiming to settle a question of the form "p or not-p?", I may, for example, decide to consult a friend knowledgeable about such

[*] I owe thanks to Yuval Avnur, Paul Hurley, and Rivka Weinberg for helpful discussion.

[†] Scripps College, Claremont, CA, USA.

[1] For a discussion of lexical considerations relating to the use of "deception" and "deceive," see Mele 2001, Chapter 1.

matters. I ask and she answers: "p." I think, "She's always right," and I come to believe that p. Alas, she's mistaken. I've caused myself to have the false belief that p and, so, I deceive myself – but not intentionally so. Deflationists about self-deception point out that such unintentional causings of ourselves to have false beliefs can frequently have a motivational or affective basis. In a simple sort of case, I may, as a result of my desire that p, spend much time thinking about p, and this may well make data in support of p very salient. I may come, in unwarranted and motivationally biased fashion, to believe that p.[2] If p is false, I've unintentionally caused myself to come to have a false belief. A fundamental issue raised in the investigation of self-deception is whether appeal to such depressingly familiar features of our cognitive lives is sufficient for the explanation of the phenomenon. Those who reject the explanatory sufficiency of such spare resources very often insist that self-deception requires more. Intentionalists insist that *real* self-deception demands that a subject *intentionally* deceive herself – that is, a self-deceiver must intentionally cause herself to come to have a false belief.

It's no doubt the case that some intentionalists find comfort in the intuition that any phenomenon worth calling "self-deception" (as distinguished from, say, "wishful thinking") must follow the contours of the processes underlying prototypical cases of interpersonal deception. Even so, it's worth noting that there's additional powerful intuitive basis for such a view. Self-deceivers very frequently believe in the teeth of the evidence and often regard as evidence for p just what the rest of us take to be – and obviously so – evidence for not-p. A self-deceiver's doxastic behavior is sometimes so striking that we are tempted to ask, "How can you possibly *believe* that?"[3] It's natural, then, to entertain the suspicion that some distinctive explanation of the self-deceiver's doxastic behavior is required. A seductive diagnosis is that self-deceivers display the light-fingered and strategic behavior characteristic of means-end rationality and, so, of intentional activity. Self-deceivers explain away just what must be explained away in order to embrace some favored proposition; they search for

[2] See Mele 1997 and 2001 for a defense of the deflationist account of self-deception and a very influential characterization of the dispute between the intentionalist and deflationist.

[3] Of course, the very nature of the phenomenon, self-deception, is increasingly disputed. In particular, it has been denied that those we describe as "self-deceived" believe what they are self-deceived about. See, for example, Gendler 2007 and Elga 2009. Here I'll just take it for granted that those who are self-deceived *believe* what we take them to be self-deceived about.

evidence favoring their focal hypotheses; they do not consider just what must not be considered. Self-deceivers, it seems, aren't trying to settle a question of the form "p or not-p?"[4] Rather, they are trying to come to believe a particular proposition. They intend to deceive themselves – they try to cause themselves to hold a false belief or try to come to believe that p regardless of the truth of p.[5] Moreover, it seems, they sometimes succeed in coming to believe that p and, so, succeed in intentionally deceiving themselves. The powerful suspicion, then, is that self-deceivers intelligently and intentionally direct, control or guide their belief-forming processes in ways that truth-oriented (or at least, non-self-deceptive) hypothesis testers do not. Their belief-forming processes are sensitive and responsive to their practical (and non-epistemic) aims in the way our intentional behavior is, generally, so sensitive and responsive. Only something like this could explain the remarkable doxastic behavior of self-deceivers. That, at least, is the intuition.[6]

[4] One way of characterizing the different aims of the intentional self-deceiver and the normal hypothesis tester is to note that the self-conscious inquirer, in aiming to settle a question, turns to the world, seeking considerations that bear on her question. An upshot of this is that, if things go well, her evidentiary or reasons condition will become determinative for belief is the following sense:
She'll come to believe that p, if by her then current lights she has sufficient reason to believe that p; or
She'll come to believe that not-p, if by her then current lights she has sufficient reason to believe that not-p.
In self-deception, as imagined by the intentionalist, things are different. The self-deceiver isn't interested in settling a question. She doesn't aim to turn to the world to seek considerations that bear on her question. Her explicit aim is precisely *not* to be doxastically open to alternatives (1) and (2). In this way, the aims or intentions of the putative intentional self-deceiver and the subject engaged in settling a question are at odds with each other. They are inconsistent aims.

[5] William Talbott (1994) characterizes the goal of self-deception so: «It [...] involves intentionally biasing one's cognitive processes to favor belief in p, due to a desire to believe that p regardless of whether p is true» (p. 30). Talbott rejects a contradictory beliefs requirement. In rejecting such a requirement, he aims, thereby to avoid a strong "divisionist" or partitioning account of the self in self-deception (p. 29). Jose Bermúdez makes note of three distinct ways – «in ascending order of strength» – in which we might characterize "core episodes" of self-deception à la intentionalism: (1) as involving «the intention to bring it about that one acquires a certain belief»; (2) as involving «holding a belief to be false and yet intending to bring it about that one acquires that belief»; and, (3), as involving «intending to bring it about that one acquires a false belief» (2000, p. 310). Nothing I say hinges on a contradictory belief requirement.

[6] This is, I think, one way of putting the perennial appeal of traditional accounts of self-deception. In familiar fashion, a traditionalist about self-deception will hold of a self-deceiver that:

1) He believes some proposition, not-p – or believes that, given the evidence, he ought to believe that not-p).
2) He engages in intentional activity the aim of which is his acquisition of the belief that p.

In this paper I focus a critical eye directly upon intentionalism about self-deception. Needless to say, intentionalism has been the object of intensive critical scrutiny by skeptics (Haight, 1980) and by deflationists (Mele, 2001) about self-deception. Much of this work has focused upon the difficulties implicated in the effort to carry out the intention to deceive oneself.[7] I am less interested in the allegedly self-defeating nature of such an effort *per se*, than I am in trying to get a grip upon the nature of the control over their belief-forming processes that self-deceivers, à la intentionalist accounts, exercise. The intentionalist holds that a self-deceiver

i. intentionally deceives herself; that is, she
ii. intentionally causes herself to come to have a false belief. [8]

Thus, the self-deceiver exercises the sort of control over her belief-forming processes that, in standard cases of interpersonal deception, the deceiver exercises over the deceived's belief forming processes and that we, more generally exercise in intentionally altering states of affairs in the broader world in non-basic action. This is the distinctive form of control over her beliefs that the self-deceiver exercises.

I'll argue here that interpersonal deception is not an available model for the sort of allegedly distinctive control the self-deceiver exercises over her belief-forming processes and beliefs. Such a view can seem plausible only by failing to recognize the real limits on our capacity to exert intentional or agential control

3) He believes, at least for a time, both that not-p and that p.

In a much cited passage, Donald Davidson appears to have embraced these three elements of a traditionalist conception of the phenomenon; he puts it so: «The acquisition of a belief will make for self-deception only under the following conditions: A has evidence on the basis of which he believes that p is more apt to be true than its negation; the thought that p, or the thought that he ought rationally to believe that p, motivates A to act in such a way as to cause himself to believe the negation of p. The action involved may be no more than an intentional turning away from the evidence in favor of p, or it may involve the active search for evidence against p. All that self-deception demands of the action is that the motive originate in a belief that p is true [...] and that the action be performed with the intention of producing belief in the negation of p. Finally, and this is what makes self-deception a problem, the state that motivates the self-deception and the state that produces it co-exist.» (1985, p. 145)

[7] See, for example, Alfred Mele's discussion of the "dynamic puzzle" (2001, pp. 13-14).

[8] The deflationist holds that self-deception is, rather, a matter of a subject

iii. non- or unintentionally deceiving herself; that is, she
iv. non- or unintentionally causes herself to come to have a false belief.

over our doxastic states.[9] In this essay, I concentrate attention on a kind of case in which an agent allegedly intentionally causes herself to come to have a false belief – a kind of case that has long-figured in discussions of self-deception but whose significance, if I am right, has not been fully appreciated. I hope to show that contrary to appearances, the agents in such cases do not intentionally cause themselves to have false beliefs – do not intentionally deceive themselves. Indeed, if we take the model of interpersonal intentional deception seriously, we ought to conclude that a self-deceiver, so regarded, deceives herself *unintentionally*. I conclude that the failure of intentionalism – or at least an intentionalism that looks to the sort of control a deceiver exercises over the deceived in interpersonal deception – in such cases constitutes indirect support for deflationist accounts of self-deception.

2. Intentional Self-Deception?

An instance of the sort of case I have in mind is this:

Happy Days : Sammy is a talented, youngish mathematician. Since his youth he's devoted himself to his career and he has enjoyed some not inconsiderable professional success and acclaim. Still, his devotion to mathematics has taken a toll on other areas of his life. He has no real friends, no lovers, no hobbies or other avocations. Sammy knows that colleagues and acquaintances derive great satisfaction from these things. He understands that there is joy attached to human intimacy but, he thinks, so long as he can do and appreciate good mathematics, he is satisfied, indeed delighted, with the trajectory of his life. Even so, there is a problem: Sammy knows that one's ability to do creative and original mathematics ebbs dramatically as one ages. Worse, still, is the fact that Sammy's family has a depressingly systematic history of early on-set Alzheimer's disease. So, not only is there reason to believe that at a certain point in his life he will be unable to gain satisfaction from the pursuit of mathematics, there is reason for believing that he will be unable to reflect backwards upon his past achievements or to take delight in the work of younger mathematicians. Sammy does believe, however, that he might

[9] My argument here, it's worth noting, is, by my lights, a development of a suggestion made by Jon Elster that beliefs are instances of states that are essentially "by-products" – states that cannot be «brought about intelligently and intentionally» (1983, p. 43). Elster also notes that such states are such as to resist or thwart «indirect as well as direct attempts to bring them about» (1983, p. 57).

well gain satisfaction, even after the on-set of illness, from reflecting backwards on a life devoted to less intellectually demanding pursuits. Of course, Sammy might change his ways now and seek out human companionship and intimacy. But why should he? The pursuit of mathematics is what now offers him the greatest satisfaction. It is far from obvious that discounting the future in this way is irrational. It seems as if Sammy is leading his life up a blind alley. But there is a solution. He now embarks upon a complex strategy designed to bring it about that he come, later in life, to believe that he has led a life rich in human connections. He fills many notebooks detailing imagined friendships, love-affairs and travels. He offers a bounty to those he engages via social media who send photographs, postcards, and letters, and other memorabilia detailing imagined intimacies with him. He secures the services of a trustee who will make certain that the relevant materials are delivered when likely to prove effective. There's no real barrier to our imagining that this strategy could succeed in the way Sammy foresees. We can imagine that, many years later, as he sits in bed at an Alzheimer's center, he's asked by an inquisitive volunteer if he has many friends or has traveled to exotic places. Sammy may say "I don't remember." Seeing the many boxes marked "friends" and "travels," the volunteer may ask, "Perhaps we should look in those? And Sammy may reply, "Yes, let's do that." He is delighted to discover that, as he now comes to believe, he has led a life that touched (and was touched by) so many others.

Such cases have been regarded by some as obvious cases of intentionally causing oneself to come to have a false belief, by others as obviously *not* cases of self-deception and by, still, others as unclear cases of self-deception. Mark Johnston (1988), Alfred Mele (2001), Brian McLaughlin (1988) and Donald Davidson (1985) all consider structurally similar cases.

Davidson writes of such a case that it «is not a pure case of self-deception, since the intended belief is not *sustained* by the intention that produced it, and there is not necessarily anything irrational about it» (1985, p. 145, n. 5). The chief source of Davidson's worry about counting such a case as a case of self-deception is his conviction that robust self-deception involves a continuing form of internal irrationality that requires the subject, a least for a time, to have contradictory beliefs. As he puts it, «the state that motivates the self-deception and the state that produces it co-exist» (1985, p. 145). Since I am concerned here solely with the demand of the intentionalist that self-deception requires that the agent intentionally cause herself to have a false belief, I cannot rely upon this sort of worry.

Brian McLaughlin (1996, p. 41) notes that one basis for holding that such cases do not count as self-deception (but are, rather, instances of what he usefully terms «self-induced deception») is that, e.g., Sammy's belief, given his evidence, is not epistemically unwarranted but being self-deceived with respect to p requires that one's belief that p be epistemically unwarranted.[10] While I agree that this is a symptom of the fact that that Sammy's isn't a case of intentional self-deception, I can imagine an interlocutor insisting that this is, rather, a mark of *really* successful intentional self-deception. After all, in successful cases of interpersonal deception, the belief the deceived individual comes to have is typically warranted. Needless to say, this reply is all the more plausible if we jettison the contradictory beliefs requirement for self-deception.

While Mele notes that such cases are «remote» from «garden variety self-deception» (2001, p. 16), he does conclude that such cases make clear that «[i]ntentionally deceiving oneself is unproblematically possible» (2001, p. 16). After all, if intentional deception is a matter of intentionally causing a subject to believe what is false then, e.g., Sammy's causing himself to believe what is false seems no less intentional than if, say, he had perpetrated the ruse on his aged father. Sammy has a plan for bringing it about that he comes to believe as he does. Events transpire as he foresees. Surely, in such circumstances he intentionally deceives himself – intentionally causes himself to come to have a false belief.

So, even if, as Mele rightly notes, Sammy's case is very different from typical cases of self-deception, such cases apparently display the fact that self-deception *can* be modeled on interpersonal deception. And this is a fact – if it is a fact – that the intentionalist might hope to exploit.[11]

Mark Johnston makes dialectical use of such cases: his aim is to show that cases like Sammy's make essential use of an «autonomous means» – a means to an end the operation of which does not require and, sometimes, does not permit agential attention to them «under the description «means of producing

[10] In this regard, it's worth noting that Sammy in *Happy Days* would appear not to satisfy Mele's "impartial observer test" for self-deception. See Mele 2000 (pp. 106-110) and 2003 (p. 164).

[11] For example, Bermúdez (2000) might well be understood to exploit this fact in his defense of intentionalism.

in me the desired belief'» (1988, p. 77).[12] This is in aid of showing that cases of self-deception that do *not* involve such means involve sub-intentional mechanisms rather than intentional activity. In his consideration of cases like *Happy Days*, Johnston writes:

> [I]t is important that one does not intend or monitor the process throughout. But, then, the operation of the means, though intended to occur, is not an intentional act and neither is the outcome produced by the means, although it is an intended outcome of a means one set in motion. [...] One intended to deceive oneself by arranging misleading evidence and taking the amnestic drug. But what one did in arranging the evidence and taking the amnestic drug did not itself constitute self-deception. Only the cooperation of future events made what one did deserve the name of deceiving oneself by arranging misleading evidence and taking the drug. So: [...] nothing that itself constitutes motivated believing or motivated cessation of (conscious) belief is an intentional act. In cases of self-deception and repression in which autonomous means are employed, the motivated believing and accompanying repression are constituted by the intentional acts of setting the means in motion plus the brute operation of the means culminating in the belief and the forgetting. [...] Even where there is a self-deceptive or repressive action plan, no intentional act is intrinsically a self-deception or a forgetting. (1988, p. 78)

I'm in sympathy with these remarks. Still, if we are modeling self-deception on interpersonal deception, it is not apparent why Sammy's actions (generating the false evidence, arranging to have it delivered) are any less "intrinsically" (or otherwise) acts of intentional deception than various acts that constitute interpersonal deception. In interpersonal deception, in the simplest sort of case, if there is an act that *is* an act of deception, it is presumably my act of saying to you that q (when we both regard q→p to be obvious), with the aim of getting you to believe falsely that p. Issues of causal deviance aside, if my act causes you to come to believe that p, I've intentionally deceived you. If to intentionally deceive is to intentionally *cause* another (or oneself) to believe

[12] In Sammy's case there are various autonomous means: his anticipated cognitive decline together with his arranging of the materials to be delivered to him at the appropriate time, etc. Autonomous means figure in various practical contexts, of course. The Soviets' doomsday device in Stanley Kubrick's *Dr. Strangelove* is an autonomous means. Autonomous means, in more familiar contexts, operate to produce ends in the face of, for example, the anticipated failure of attention or a short-term change of preference.

what is false, then, the act which is intentionally performed (with an eye to producing false belief) is the act which is the act of deception.[13]

The mere fact that Sammy (and others like him) no longer consciously intends to deceive himself for some period of time prior to coming to believe that p should, by itself, be no obstacle to our viewing Sammy as intentionally deceiving himself. Certainly, in a case of interpersonal deception, as with other such cases of non-basic actions, once I do whatever I do – for example, assert that q – to initiate the casual chain that results in your coming to believe that p, my contribution is over. I need no longer actively intend or monitor the situation. Indeed, as deceiver I could *die* during the temporal interstice between my act and the deceived's coming to believe and, yet, I would, nonetheless, count as having deceived you.[14] In familiar cases of non-basic action, say, sinking a putt, my contribution is over – body English aside – once I strike the ball. Yet I sink the putt, if acting as I do, I cause the ball to drop into the cup. So it can't be the fact that, in Sammy's case, there's a point at which he can't or doesn't intervene in his deception that makes it the case that his self-deception is not intentional.

So, should we conclude that *Happy Days* and other similar cases are cases of intentional self-deception? We should resist such a conclusion. In this regard, we do well, I think, to ask how a subject might *try to bring it about that he unintentionally deceives himself that p.* (We can imagine that something important – a large wager or his life – hinges upon his coming to believe that p and upon his doing so in unintentional fashion.) It seems to me that he might do this via an effort to arrange evidence in such a way that, at some later point, he comes to believe that p and that he does so as a result of his, then, good-faith effort to settle the question "p or not-p?" If this is so, Sammy's effort to deceive himself intentionally and our current subject's effort to unintentionally deceive himself look to be no different.

It might, I suppose, be suggested that someone who aims to bring it about that he non- or unintentionally deceives himself that p must resort to other sorts of maneuvers. Perhaps, what such a subject must do is, e.g., to seek out experts on p-related matters and simply ask "p or not-p?", believing that they are experts but *hoping*, somehow, that they will offer erroneous counsel. In

[13] Presumably, whatever we mean by an act that *is* an act of self-deception we cannot mean an act that is somehow constituted by the *coming* to believe what is false.

[14] Such a view is not mandatory, of course; see Sorensen 1985.

such circumstances, if an expert says "p" and the subject, believing the expert is always right, comes to believe that p, she'll have deceived herself via her asking the expert.[15] But this hardly seems like a way of *trying* to deceive oneself unintentionally. Indeed, such a "plan" for bringing it about that one unintentionally deceives oneself seems no different than trying to settle the question "p or not-p?" but hoping, somehow, that one gets it wrong.

But if this is right, then, we seem to be in a position of concluding either that what one does when one's trying to intentionally deceive oneself and what one does when one's trying to unintentionally deceive oneself are no different or, perhaps, worse, that when one intentionally deceives oneself one also unintentionally deceives oneself. Needless to say, it may be claimed that the effort to unintentionally deceive myself is, somehow, essentially self-defeating. There would, of course, be an irony here since we've become used to thinking that it's, rather, the effort to bring it about that I intentionally deceive myself that is essentially self-defeating.

3. Who Deceives Whom?

Since we are considering a potentially puzzling consequence of the effort to model intentional self-deception on prototypical cases of interpersonal deception we would do well to consider, in brief, the nature of the activity of the deceiver and the deceived in interpersonal deception. If to deceive another is to cause her to believe falsely that p, we should be clear that what the deceiver's action causes is an event - presumably the event of the deceived's coming to believe that p. I take it that this is so in virtue of the fact that a deceiver alters the evidence or epistemic reasons of the deceived and this results in the latter's coming to believe that p. In this way, if all goes well (for the deceiver, that is), the deceived's belief-forming processes are sensitive and responsive to the deceiver's intentions and practical reasons in the way that the path of the ball is sensitive to my aims when I sink a putt.[16] In this way, we are

[15] Thus, the expert, as well, will have unintentionally deceived the subject.

[16] It's important that the control I exercise over the deceived in cases of intentional deception is not merely causal. Consider the following: I assert that p to you, believing it false and thinking you regard my testimony as trustworthy. As it happens, you aren't at all inclined to believe on the basis of my assertion alone. Still, you've just emerged from a session with your much esteemed psychic. You've consulted him, as you're consumed with the question "p or not-p?" since you desperately desire that

right to regard deceiving another as treating another as a mechanism. In familiar fashion, we exploit machinery and the causal structure of the world, more generally, in order to extend the range of our control and, so, to secure our ends. In this way the deceiver acts upon and through the deceived.

I take it that Iago is remarkably successful in this regard with respect to Othello in the matter of the question of Desdemona's fidelity. Iago, in this sense, treats Othello as a mechanism in order to secure his aims. He exercises control over Othello's reasons and belief-forming process. And that is why we say he deceives Othello – causes Othello to come to believe that Desdemona is unfaithful. Iago accomplishes his deception via the alteration of Othello's evidentiary or reasons condition. Believing p false and wanting Othello to come to believe that p, Iago arranges things such that Othello comes to possess evidence in favor of p; his reasons condition becomes determinative for p and he comes to believe it. Iago intentionally deceives Othello – causes him to believe something false. Presumably, this is something Iago does. One agent intentionally deceives another via the first agent's pursuit of a deceptive project that exploits the second agent's pursuit of the project of settling a question. So, there are two projects simultaneously at play – two projects traceable to two agents and to two constellations of practical reasons.

This is, of course, one reason why it is nonsense for a deceiver to say to a deceived: "Don't look at me. Your coming to believe that p is something *you* did, not *me*. You came to believe for your own reasons." This is nonsense even though Othello's coming to believe or forming the belief that p is not something *Iago* does. Othello *does* that and for his reasons. In this way, Othello is not a passive by-stander to his deception. But this should be no surprise. Deceptive projects in the interpersonal arena exploit the rational

not-p. He's just told you that if you can get to midnight without hearing a typically trustworthy speaker assert that p, you can be assured that not-p is true – otherwise, p is true. You immediately try to make your way home to seek seclusion, when you encounter me. Now, my assertion certainly plays a causal role in your coming to believe that p; yet, if p is false, I don't intentionally deceive you. We have a case of consequential waywardness or deviance. But it's important to point out that this is so because what I do fails to exert the sort of control that I aim to exercise over the direction your cognition. My intention to cause you to believe that p is, of course, not appropriately related to how it is you are caused or come to believe that p. Here it seems to me, were I to come to learn why it is you came to believe that p, I might well reasonably say: "Your coming to believe that p is something you did or brought about, not me!" In such a case, the deceiver may certainly be said to cause the deceived to come to believe as he does, but he does not intentionally deceive. I lack the appropriate sort of control over your being caused to come to believe as you do.

agency of another. Iago is trying to deceive Othello. Othello is trying to settle a question. Othello (with the assistance of Iago, to be sure) takes certain data to constitute powerful evidence in favor of the view that Desdemona is unfaithful. In focusing upon various data he causes himself to come to believe this. So he causes himself to have a false belief. In this way, Othello deceives himself – but unintentionally, of course – and in the manner that we all often deceive ourselves in unintentional fashion. Unless Iago could somehow directly implant the belief that Desdemona is unfaithful into Othello, it's hard to see how this result is avoidable.

Iago deceives Othello. But he does something else: he intentionally causes Othello to deceive himself unintentionally. Thus, typical cases of interpersonal deception require the presence of intentional deception (on the part of the deceiver) and unintentional deception (on the part of the deceived). With this as a model for intentional self-deception, we may want to say of Sammy that he:

(1) He intentionally causes himself to (come to) have a false belief; that is,
(2) He intentionally deceives himself.

But, as well, when Sammy comes to believe as he does, he does so in the aftermath of his effort to settle a question. He takes various data to constitute sufficient reason for settling his question. In this way Sammy, like Othello,

(3) Unintentionally causes himself to (come to) have a false belief; so, he
(4) Unintentionally deceives himself.

And this, of course, because Sammy, like Iago in his deception of Othello,

(5) Intentionally causes himself to deceive himself unintentionally.

This is the source of the puzzle at the end of the last section. If interpersonal deception is our preferred model, then we must conclude that if Sammy were to aim to deceive himself *unintentionally* he could do no better than to do precisely as he does in the case as described in *Happy Days* – a case in which he, of course, allegedly intentionally deceives himself; and this, because, as we now see, Sammy *does* unintentionally deceive himself in that case. Indeed, there is an additional puzzle since Sammy both, (2), intentionally deceives himself and, (4), unintentionally deceives himself. Thus, the very same doxastic alteration, at the very same time, by the very same agent must be counted an instance of both intentional deception and unintentional deception. Of course, we might insist that if Sammy in *Happy Days* unintentionally deceives himself he does not, as well, intentionally deceive himself.

It is precisely because of the presence of two agents with two distinct projects in cases of familiar interpersonal deception that there is no puzzle attached to conceiving of Iago's deception of Othello as involving both intentional and unintentional deception – and this, of course, because Iago intentionally deceives Othello, while Othello deceives himself unintentionally. So, the presence of two agents with two distinct projects is crucial to our understanding of interpersonal deception and, in particular to the way in which one agent may intentionally cause another to come to believe falsely that p and, in this way, to control or manipulate the belief-forming processes of another agent via her (i.e., the deceiver's) deceptive intentions and intentional activities. The deceptive agent counts upon or exploits the fact that the deceived is engaged in and pursuing her own project: settling a question or trying to get things right. Iago intentionally causes Othello to come to have a false belief via his pursuit of his deceptive project. Othello deceives himself unintentionally via his effort to settle a question. So, again, on this interpersonal model, we say of Sammy that he intentionally causes himself to have a false belief via his pursuit of his deceptive project while he also unintentionally deceives himself via his effort to settle a question. At the least, we're compelled to view Sammy as possesses two competing and contrary projects.

But there's just one Sammy. Now, this might be disputed, of course. In obvious ways we can claim that it is the earlier time-slice of Sammy who succeeds in intentionally deceiving the later time-slice of Sammy, while the later time-slice of Sammy unintentionally deceives himself in the midst of his trying to settle a question. To be clear, though, Sammy comes to believe that p at a particular time; so, at that time Sammy's deceptive project succeeds *and* Sammy unintentionally deceives himself. But this is to treat Sammy not merely as if he were like two distinct agents but, rather, as if he were, in fact, two distinct agents. And the cost here, it seems to me, is very great.

If self-deception literally implicated two agents or two independent centers of rational activity, I take it that it would be foolhardy to gainsay the possibility of intentional self-deception. Needless to say, there are accounts on offer that appear to involve something like this strategy (Pears, 1984; Rorty, 1988). Still, I take it that there's something profoundly unsatisfying about such radical homuncular accounts. If, we explicate intentional self-deception by appeal to two independent centers of rational activity, we will have failed to come to grips with the phenomenon and what we find puzzling about it. We would have failed to come to grips with the phenomenon because we would have turned a case of self-deception into a case of interpersonal deception. And we would have failed to explain what we find puzzling about the phenomenon ("How can you possibly *believe* that?") since there's nothing puzzling about how or why one comes to believe as a result of the activity of a deceiver. (At best we would have explained away our puzzlement.) Rather, my point is that if self-deception, à la intentionalism, is to be compellingly defended and explained, the phenomenon must be realized, as William Talbott puts it in a single, coherent self (1996). If what we call "self-deception" involved one center of rational activity or agent controlling the epistemic reasons possessed by another independent center of rational activity in precisely the way Iago controls Othello's reasons, there is, it seems to me a straightforward way in which we would have to conclude that there is no self-deception.

What should we say about Sammy in *Happy Days*? I think we should say, (5), that Sammy intentionally causes himself to deceive himself unintentionally,[17] but that we should resist saying that he intentionally deceives himself. Sammy tries to bring it about that he unintentionally deceives himself. He does unintentionally deceive himself. Of course, one imagines the immediate rejoinder: but then he also must, (2), intentionally deceive himself. If he intentionally causes himself to unintentionally deceive himself *then* he intentionally deceives himself. Indeed, Sammy, we will say, *intentionally* deceives himself by *unintentionally* deceiving himself.[18] My own view is that we can say this only if Sammy is treated precisely like Iago and Othello – as two distinct agents with two distinct projects and constellations of practical reasons. In the next section, I aim to consider why, in the case at hand, we should reject the suggestion that, in these circumstances, Sammy intentionally deceives himself by unintentionally deceiving himself.

[17] More felicitously we can say that Sammy intentionally brings about conditions in which he unintentionally deceives himself.

[18] Needless to say, an agent can intentionally ϕ by unintentionally ψ-ing. For example, I can intentionally amuse the children by intentionally causing myself, unintentionally, to trip down the stairs. But in this case, the intentional causing (an action) produces my unintentional tripping which then produces a distinct event: the children's merriment. In the case of self-deception, though, it is the causing to believe what is false that is both intentionally and unintentionally produced.

Before turning to that task, I want to note that the modest conclusion that Sammy intentionally deceives himself via unintentionally deceiving himself would, itself, appear to have awkward consequences for intentionalists. For while it may be insisted that it is clear how, in Sammy's case, intentional self-deception succeeds, it is far from clear how, without similar improbable contrivances (e.g., Alzheimer's-induced forgetfulness together with fabricated, but compelling, evidence delivered by a trustee, etc.) intentional self-deception could succeed. Indeed, as Mele notes, such cases as *Happy Days* are remote from typical cases of self-deception; and they are so in part precisely by virtue of the presence of such fanciful elements. Those fanciful elements are, of course, critical to Sammy's coming to believe as he does. He comes to believe as he does, in the midst of settling a question because he comes to have sufficient reason so to believe. But, then, we must ask, how without such contrivances is intentional self-deception to succeed?

Here it should be pointed out that instances of intentional self-deception either involve intentionally causing oneself unintentionally to deceive oneself or they do not. If they do, then, in the absence of the baroque elements critical to success in *Happy Days* some other mechanisms and processes must be at work which result in a subject's unintentionally deceiving herself. If success hinges upon intentionally causing myself to deceive myself unintentionally, it is not at all easy to see what these other mechanisms and processes could be if not the non-intentional motivational and affective mechanisms described by deflationists. After all, the self-deceiver must be moved to regard her data as sufficient reason for belief.

Of course, it may be that the intentional self-deceiver does not succeed in intentionally deceiving herself by unintentionally deceiving herself while in the midst of trying to settle a question. That is, it may be that there are not two projects – the deceptive project and the effort to settle a question – at work. A natural way of developing this suggestion is to appeal to unconscious deceptive intentions and projects (Talbott, 1994; Bermúdez, 2000). While it is certainly the case that I cannot take up this challenge with the attention it deserves, one consequence of this view should be noted. Appeals to unconscious deceptive projects and intentions are very often accompanied by an insistence that the requisite sort of unconscious is a familiar one – an innocent or minimal unconscious (Talbott, 1994; Bermúdez, 2000). William Talbott insists, for example, that the sort of unconscious upon which his account relies «requires no more division of the self then does explaining ordinary communication, or explaining such activities as singing a duet, or painting a house together [...]» (1994, pp. 36-37).[19]

Thus, the claim is that in intentional self-deception there are not two competing projects or intentions, there is just one: the self-deceptive project of intending to come to believe that p (regardless of its truth.) Still, there is the stubborn fact that self-deceivers — in the midst of deceiving themselves — do take themselves to be doing whatever they are doing when they, in fact, are trying to settle a question. So, at the very least, in such cases, an agent who intends to deceive herself, and whose activities through time are presumably organized and directed toward that end, also takes herself to be settling a question. Moreover these are projects or intentions that are at odds with each other. On such a view, the agent isn't merely ignorant of the project she's really engaged in and of the intentions and reasons animating it; she is positively mistaken about what she is doing; in particular, she is mistaken about why, when, for example, she rejects a datum as probative, she is rejecting that datum as relevant. Such an agent takes herself to be trying to settle a question, takes herself to be organizing her activities toward that end, but she is not. She is, in fact, engaged in the contrary project of trying to deceive herself. But this seems less a familiar and innocent appeal to an unconscious of the sort present in communicative activity or to "innocent" divisions of the self, than it does an appeal to a robustly psychodynamic conception according to which our conscious projects and aims are epiphenomena floating powerlessly above of our unconscious intentions, aims, and reasons.

4. Occluded Reasons

The challenge to which I now return is this: to intentionally deceive is to intentionally cause to believe falsely. Sammy, I have suggested, intentionally causes himself to deceive himself unintentionally. That is to say (rebarbatively): Sammy intentionally causes himself to cause himself unintentionally to come to have a false belief; or (somewhat less rebarbatively), Sammy intentionally brings about conditions in which he unintentionally deceives himself. But, again, if Sammy intentionally causes himself to deceive himself unintentionally,

[19] Talbott appeals to Grice on communicative intentions and to the intentions that figure in Bratman's theory of shared or joint activity as analogues to the unconscious intentions implicated in self-deception.

then it seems that he intentionally deceives himself (*by* getting himself to deceive himself unintentionally). Moreover, the same conclusion seems to result when we make note of the fact (apparent in the rebarbative formulation above) that an intentional causing of a causing surely collapses into an intentional causing; that is, if Sammy intentionally causes himself to cause himself unintentionally to come to believe falsely that p, then, he intentionally deceives himself.

Why, then, deny that Sammy (and others) intentionally deceives himself in circumstances in which he intentionally causes himself to deceive himself unintentionally? I will argue – too briefly here – that Sammy's earlier intention and practical reasons are occluded or screened off from playing an intentional or rationalizing explanatory role in his deceiving of himself.

To see how this is so, consider a case, from the strictly practical sphere, described by Alfred Mele. In the case, Ann is offered $10,000 if she offends Bob unintentionally. «Ann,» Mele writes

> will be inclined, in some measure to bring it about that she offends Bob unintentionally. In one relevant scenario, she knows that she tends to offend Bob unintentionally when she is extremely busy: when she is preoccupied with her work, for example, she tends, without then realizing it, to speak more tersely than she ordinarily does to people who phone her at the office; and, when Bob calls, her terse speech tends to offend him. Knowing this, Ann may undertake an engrossing project [...] with the hope that her involvement in it will render her telephone conversation at the office sufficiently terse that should Bob call (as he frequently does), she will unintentionally offend him. This is a coherent attempt [...]. (1995, p. 414)

That seems right. When Ann offends Bob by speaking tersely to him that evening, she does so for considerations then salient to her and not in virtue of the considerations salient to her when she formulated her plan. She acts thoughtlessly and unintentionally. She does not offend Bob intelligently and intentionally. What about the fact or state affairs <Bob's being unintentionally offended by her>? She does intentionally bring about or cause that *state of affairs*; but this is to say that she intentionally brings about conditions in which she insults Bob unintentionally. And this is consistent, of course, with her exerting no intentional control or guidance via her earlier practical reasons over her current treatment of Bob. Luckily for Ann, those have come to be screened off by the interposition of her current motivational and cognitive constitution. Of course, in the aftermath of her success, Ann may think:

"Yahoo! I've done just what I wanted to do – the $10,000 is mine." But for all that, she does not exert (in virtue of her practical reasons at the time she formulated the plan) intentional control over her offending of Bob. At the time she formulates her plan, she foresees that she will offend Bob but that, too, is consistent with the claim that she unintentionally offends Bob when she does. Of course, it's clear that what she has done at an earlier time as well the practical reasons then animating her activities are causally relevant to her later unintentional offending of Bob. But it is not in virtue of those that she does what she does when she offends Bob.

How, then, does Ann succeed in bringing it about that she unintentionally offends Bob? Well, what she must do is to arrange things such that she will come to have a different constellation of reasons and a different aim from those that give rise to the original aim or project. It is, of course, the temporally later set of reasons that produces her action whereby the state of affairs <Bob's being unintentionally offended by her> is realized – the state of affairs that Ann aimed to bring about, given her earlier reasons. In short, what she does when she offends Bob is explained by the reasons she has come to acquire: she's working hard in the evening, doesn't have time for a meandering conversation and wants to get off the phone. She answers Bob's question tersely wanting to get off the phone and he is thereby offended. The reasons from which her earlier aim (i.e., offending Bob unintentionally) emerged explain – in the rationalizing way – not why she acts as she does when she offends Bob, but rather why she comes to have the reasons that, at the later time, explain her acting as she does. So, while it's certainly the case that her earlier reasons and intention figure in the causal explanation of her later activity, they do not figure in the intentional or rationalizing explanation of her later activity. What she does then is explained by the reasons she has come to possess at the later time. What's crucial here, again, is that the practical reasons which generate her project are screened off – in ways she hopes will occur – from those which generate her later behavior.

Thus, in Ann's case, we will say that she intentionally causes herself to insult Bob unintentionally.[20] Let me be clear about the relationship of this case to that of intentionally deceiving oneself: since to deceive oneself is to *cause* oneself to have a false belief, the structural analogue in the case of alleged

[20] Or we may say, a bit more felicitously, that she intentionally brings about conditions in which she unintentionally insults Bob.

intentional self-deception is this: Sammy intentionally causes himself to cause himself unintentionally to come to hold a false belief. In more familiar settings, the intentional causing of a causing will collapse into an intentional causing. But not so in these cases, since the means by which one brings about the state of affairs one wants to bring about entails that one's reasons-condition and intention be altered in order that the desired state of affairs be the upshot of a distinct reasons condition and intention.

Sammy comes to believe as he does because he's motivated to settle a question and he comes to settle his question in virtue of the epistemic reasons he comes to possess – that is what explains his coming to believe as he does. But, as well, his causing himself to come to have a false belief is something explained by his then current aim and reasons. In the midst of settling a question, he asks to see what's in the boxes and, as a result, comes to believe that he's led a life rich in human connections, and, he thereby deceives himself unintentionally. His earlier reasons are, like Ann's, occluded or screened off from providing a rationalizing account of his deception of himself. Something like this point is noted by Jonathan Bennett; he argues that agents can be appropriately said to act through «long, complicated causal chains but not ones whose whole effectiveness runs through the will of an agent» (Bennett, 1988, p. 227). He writes that

> at noon I set up a delayed-action mechanism, knowing that when it kicks into action at dusk it will irresistibly tempt me to close the gate. In that case, what qualifies me as the one who closes the gate is what I do at dusk not what I do at noon. (Bennett, 1988, p. 227).

Sammy at the time he comes to believe he has led a life rich with human intimacy comes to believe as he does as a result of his effort to settle a question and the epistemic reasons he comes, then, to possess. In this way, his earlier plan and intention to bring it about that he deceives himself is one whose, as Bennett puts it, "whole effectiveness runs through the will of an agent."[21] The parallel between Bennett's gate-closer and *Happy Days* case might appear to be vitiated by the fact that Bennett does, of course and rightly, want to speak of

[21] I am certainly not presuming that there is a "doxastic will." I am presuming that trying to settle a question is an intentional activity and, as well, that settling a question – i.e., coming to believe, as it may be, that p – is an instance of rational activity. That I come to believe as I do is something I do because of my apprehension of reasons. See, for example Raz 1999 (Chapter 1) and Moran 2002.

the agent in this case as (by virtue of what he does at noon) causing himself to close the gate at dusk (1988, p. 227). But, of course, Sammy is trying to deceive himself, which is just to try to *cause* himself to come to have a false belief. So, I agree that Sammy intentionally causes himself to deceive himself. This is what I have been arguing: Sammy does not intentionally cause himself to come to have a false belief – what he does is to intentionally cause himself unintentionally to deceive himself. Less awkwardly, he intentionally brings about conditions in which he unintentionally causes himself to have a false belief.

Thus, when I intentionally cause my own action, when that action is *already* a causing, as with deception, then, after Bennett, we should say that Sammy counts as deceiving himself in virtue of what he does while in his bed at the Alzheimer's center, rather than in virtue of what he does as a young mathematician. As with Ann, his earlier reasons and intention are occluded or screened off from providing a rationalizing explanation of his deceiving of himself. In this case, then, the intentional causing of an unintentional deception does not collapse into an intentional deception and this because, like Ann, Sammy now has another aim and constellation of reasons, and these provide the rationale for his coming to believe as he does and, so, for his deception. Of course, in virtue of what he does as a young mathematician, Sammy counts as causing his later deception; but, again, this is not to say that he intentionally deceives himself – intentionally causes himself to come to have a false belief. There is no act which is an act of intentional deception.

This is why, if Sammy wanted to deceive himself unintentionally, he could do no better than to arrange things such that at some later time, while in the midst of trying to settle a question, he would come to take himself to have sufficient reason for coming to believe that p. His earlier reasons and intentions are occluded from playing a rationalizing explanatory role in his deception, as Ann's are from her offending of Bob. He comes to have a different aim, settling a question; as a result, he comes to have various epistemic reasons his apprehension of which constitutes by his lights sufficient reason for coming to believe as he does. His coming to believe as he does is explained by appeal to these propositional attitudes and by his rational activity at that time. As a result of his current activities – his inquiry – he causes himself to come to have a false belief and, so, to deceive himself unintentionally. By virtue of what he did as a young mathematician, he intentionally caused himself to deceive himself unintentionally. He does not intentionally deceive himself

5. Conclusion

I have argued that Sammy does not succeed in intentionally deceiving himself. I have, as well, pointed out that if intentional deception, in fact, requires that the agent unintentionally deceives herself, intentionalism faces serious challenges.

The interpersonal model of intentional deception is no model for self-deception because, since I am a single agent, once my evidentiary or reasons condition is altered – the condition of success of my project – I have altered the reasons condition of the actor and, in fact, have abandoned the intention to deceive prior to coming to believe. Indeed, that aim to deceive myself has been replaced by another contrary aim: the aim of settling a question. Iago's act of successful deception requires for its success the rational activity of another agent. In self-deception, there is no distinct agent to whom the deceptive project can be traced. To treat such a case as a case of intentional deception is to treat a single agent precisely as two agents. Moreover, since there is but one agent in self-deception, there is no other agent whose activity or aims could be the source of the deception. In self-consciously aiming to alter my reasons condition and my aims in order to bring it about that I come to have a false belief as a result of settling a question, I guarantee that what I do is to intentionally bring about conditions in which I cause myself unintentionally to deceive myself.

REFERENCES

Bennett, J. (1988). *Events and Their Names.* Indianapolis: Hackett.

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

Davidson, D. (1985). Deception and Division. In E. Lepore, & B.P. McLaughlin (Eds.), *Actions and Events: Perspectives on the Philosophy of Donald Davidson.* New York: Basil Blackwell, 138–148.

Elga, A. (2009). Imagination, Delusion, and Self-Deception. In T. Bayne, & J. Fernández (Eds.), *Delusion and Self-Deception.* New York: Psychology Press, 263–280.

Elster, J. (1983). *Sour Grapes*. Cambridge: Cambridge University Press.

Gendler, T. (2007). Self-Deception as Pretense. In J. Hawthorne (Ed.), *Philosophical Perspectives 21: Philosophy of Mind*. New York: Wiley Interscience, 231–258.

Haight, M. (1980). *A Study of Self-Deception*. Sussex: Harvester Press.

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

McLaughlin, B.P. (1988). Exploring the Possibility of Self-Deception in Belief. In A. Rorty, & B.P. McLaughlin (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 29–62.

Mele, A. (1995). Motivation: Essentially Motivation-Constituting Attitudes. *The Philosophical Review*, *104*(3), 387–423.

Mele, A. (1997). Understanding and Explaining Real Self-Deception. *Behavioral and Brain Sciences*, *20*(1), 127–134.

Mele, A. (2003). Emotion and Desire in Self-Deception. In A. Hatzimoysis (Ed.), *Philosophy and the Emotions*. Cambridge: Cambridge University Press, 163–179.

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

Moran, R. (2002). Frankfurt on Identification: Ambiguities of Activity in Mental Life. In S. Buss, & L. Overton (Eds.), *Contours of Agency: Essays on Themes form Harry Frankfurt*. Cambridge, Mass.: The MIT Press, 189–217.

Oksenberg Rorty, A. (1988). The Deceptive Self: Liars, Layers, Lairs. In A. Rorty, & B.P. McLaughlin (Eds.), *Perspectives on Self-Deception*. Berkeley: University of California Press, 11–28.

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

Sorensen, R. (1985). Self-Deception and Scattered Events. *Mind*, *94*(373), 64–69.

Raz, J. (1999). *Engaging Reason*. Oxford: Oxford University Press.

Talbott, W. (1995). Intentional Self-Deception in a Single Coherent Self. *Philosophy and Phenomenological Research*, *55*(1), 27–74.