(This is the author's copy of a paper published in the *Canadian Journal of Philosophy* in 2017, accessible from Taylor and Francis at http://www.tandfonline.com/loi/rcjp20.)

An Agentive Non-Intentionalist Theory of Self-Deception

Kevin Lynch

Abstract

The self-deception debate often appears polarized between those who think that self-deceivers intentionally deceive themselves ('intentionalists'), and those who think that intentional actions are not significantly involved in the production of self-deceptive beliefs at all. In this paper I develop a middle position between these views, according to which self-deceivers do end up self-deceived as a result of their own intentional actions, but where the intention these actions are done with is not an intention to deceive oneself. This account thus keeps agency at the heart of self-deception while also avoiding the paradox associated with other agency-centered views.

1. Introduction

In 2008, Sharon Collins was sentenced to 6 years in prison for conspiracy to murder in an Irish court. She was found guilty of hiring Las Vegas card-dealer Essam Eid, through his 'hitmanforhire.net' website, to kill her wealthy partner and his two sons so that she could inherit the businessman's millions. The plot was foiled when Eid went to Ireland, changed his mind about the hit, and approached the targets offering them the chance to buy out the contract, after which police were called. A lead investigator in the case said that never in his career had he seen so much evidence stacked against a person. This included detailed email conversations between Collins and Eid retrieved from computers in Ireland and the United States, phone records, records of a money transfer, a proxy-marriage certificate obtained by Collins (her partner wouldn't marry her as he wanted his sons to inherit his business), testimony from Eid's accomplice, and traces of the deadly poison ricin found in Eid's possession.

Despite the staggering amount of evidence against her, Collins' partner PJ Howard–a private man who was clearly besotted with the pretty, petite and extremely two-faced blonde 14 years his junior–refused to believe that she had tried to have him and his sons killed and sided with her story that she was framed as part of an elaborate shakedown. That he genuinely believed

she was innocent was evidenced in a number of ways. He wrote to the Director of Public Prosecutions urging him not to bring charges against her. He publicly rejected the charges against her in court when giving evidence, defended her character, and kissed her on the lips when he left the witness box. He refused to leave her, against the advice of his sons and solicitor, and visited her in jail after her imprisonment. And he spent a large sum of money hiring private investigators to verify her story for an appeal (Connolly 2008).[1]

This case demonstrates, in rather dramatic fashion, that phenomena possessing the following features exist:

A) A subject, *S*, encounters evidence warranting belief in the truth that not-*p*.
B) *S* strongly desires that *p*.
C) Because *S* desires that *p*, *S* ends up believing that *p*.

… and in light of Howard's willingness to back Collins' story with his reputation and money, there is no reason to think that he also secretly or unconsciously believed/suspected the truth. I would not be alone in thinking that this phenomenon–in short, that of *falsely believing something against significant contrary evidence because you want it to be true*–is self-deception in its most paradigmatic variety.[2] (A-C excludes 'twisted self-deception' (Mele 2001, chap.5), a non-paradigmatic form. Our focus will only be on paradigmatic self-deception here.) Of course, some believe that, paradigmatically, self-deceivers do not really believe what they want to be true but know the truth (e.g. Bach 1981). I have defended the former view of self-deception elsewhere (Lynch 2016, 518-519; 2012) and will take it for granted here.[3]

Taking this as our starting-point, the following important question arises, which will be the central concern of this article. We may call it the *explanatory question*:

---

[1] Elsewhere I have argued that a willingness to take risks on an assumption is a mark of authentic belief (Lynch 2012, 444-445.), and Howard displayed such willingness in this case.
[2] For instance, see the opening remarks in (Deweese-Boyd 2006/2012) and (Michel and Newen 2010, 731-732). This is assuming that the causal chain between the desire and belief is non-deviant, which we could expect in ordinary cases like that of PJ Howard.
[3] I have also argued elsewhere (Lynch 2012) that quite often self-deceivers may not quite fully believe that *p* (where *p* is the proposition they want to be true) but may have a degree of conviction in this proposition that exceeds what's rationally warranted while not being enough to constitute a belief that *p*. This point can be ignored for current purposes however, as it will simplify this discussion to just focus on cases where the self-deceiver believes that *p*.

How, in self-deception, does *S* end up falsely believing that *p* because of his/her desire that *p*?

That is, the subject believes that *p* against good evidence to the contrary, where she would have concluded that not-*p* were it not for her desiring that *p*. So the question arises: how does her having that desire lead to her believing that *p* against the thrust of the evidence, in ordinary real-life cases of that sort?

This question is pressing because on first glance it's puzzling how such a thing could happen. It's puzzling, because when explaining why people believe things we typically refer to the evidence/reasons they had, or thought they had, for the belief, or mention facts that suggest what the evidence was (e.g., 'he knows/believes Gdynia's in Poland because his wife's from there'). But the fact that one desires that *p* obviously does not constitute evidence at all for believing that *p*, much less evidence that could overturn the weighty evidence for not-*p*, and *S* could hardly be mistaking it for such. So how did *S* end up believing that *p* from desiring that *p*?

As I see it, approaches to answering this question in the philosophical literature have mostly been polarized between two extremes. Traditionally it was assumed that people deceive themselves in much the same way that they deceive others: by intentionally making themselves have this belief, which they know is false. Philosophers soon found this 'intentionalist' account paradoxical and took flight from it, but (as we'll see) towards an equally objectionable alternative, where self-deception is regarded as an affliction that befalls us rather than as something we do to ourselves, where desires and emotions unduly influence cognition in relatively automatic ways. The first view conceives of the self-deceiver too strongly as perpetrator, while the second conceives of him too strongly as victim. The aim here is to deliver an answer to the explanatory question that gets the subtle balance right between both the perpetrator and victim aspects of self-deception.[4] This will be achieved by giving a robust role to intentional action in the explanation of the self-deceptive belief, but where these actions are *not* done with an intention to make oneself have the belief. Self-deception *is* something that one does to oneself, but *unintentionally*.

---

[4] Attempts to reach a middle ground between these approaches have been made before, most notably by Scott-Kakures (2002). Though there is much to be admired in this attempt, Scott-Kakures endorses the 'FTL explanation' of self-deception, which, as I'll argue in the final section, is different from and inferior to the explanation offered here.

We will begin in section 2 by separating out two important distinctions that have been conflated, the intentionalism/non-intentionalism distinction, and what I call the agentivism/non-agentivism distinction. The meaning and attraction of an agentivist non-intentionalist theory is then outlined, and this perspective is then argued for in sections 3 and 4. In section 3, empirical studies are presented which show that the subject's own actions, and also omissions, are responsible for causing the self-deceptive belief. It is then argued that these actions must be intentional, and over sections 3 and 4, an account of the relevant intentions is developed alternative to construing them as intentions to deceive oneself, a task that involves outlining the rationalizing relationship between the desire that *p* and these intentions. Self-deceivers do not act with the intention of making themselves have a belief, but with other intentions, such as, primarily, the *intention to find any problems with the unwelcome evidence*, a hypothesis that, as is argued in section 4, is psychologically plausible and free of paradox. Then, in section 5, this view is related to other currently popular non-intentionalist explanations. There it is shown that this hypothesis has the power to subsume and unify many of the self-deceptive processes described in the non-intentionalist literature, thus providing a deeper explanation of the phenomenon, and it is also contrasted with the 'FTL model', a popular explanation of self-deception among non-intentionalists. The result will be an overhauled non-intentionalism which bucks the trend with such theories of explaining self-deception in terms of affective-cognitive *mechanisms*, explaining it instead in terms of the person's *rational* (in the thin sense) *actions*, that is, actions done for reasons.

2. A Misconception about Non-Intentionalism

The answer to the explanatory question on offer here is a non-intentionalist one, but before we develop this there is a widespread assumption about non-intentionalism we must be disabused of. Again, the intentionalist answer asserts that self-deceivers end up believing that *p* by intentionally making themselves have the belief, a belief they know is false/unwarranted.[5] So *non*-intentionalism is simply the *negation* of this: it is just the denial that self-deceivers intentionally make themselves have a belief that they know is false/unwarranted.

[5] There are various reasons why philosophers side with intentionalist explanations, one being the belief that self-deception must resemble other-deception. See (Bermúdez 2000) for more.

As we'll see, this leaves the non-intentionalist with a variety of explanatory options to pursue. However, the impression is often given in the literature that non-intentionalism is committed to the view that intentional action, motivated[6] by the desire that *p*, plays no role in the explanation of the self-deceptive belief. That is, it gets assumed that 'non-intentionlism' means 'no intentional action'. For instance, Bermúdez, a critic of non-intentionalism, says that '[a]ccording to anti-intentionalist accounts … self-deceiving belief formation can be explained simply in terms of motivational bias, without bringing in appeals to intentional action' (2000, 309). Similarly Talbot, a fellow critic, interprets 'anti-intentionalism' as holding that the desire causes the belief by 'triggering' or 'activating' 'non-intentional mechanisms' (1995). And Deweese-Boyd writes that 'non-intentionalist approaches … [render] the process by which [one] is self-deceived subintentional' (2006/2012).

On this way of distinguishing intentionalism and anti-intentionalism, the theoretical divide between these two persuasions appears very wide. Moreover, since the distinction is supposed to exhaust all possible answers to the explanatory question, it polarizes the debate and hampers our ability to conceive of the full range of explanatory options available to us.

However, the supposition that the self-deceptive belief exists due to the person's actions (including mental actions), actions which are intentional and motivated by the desire that *p*, *can be consistent* with non-intentionalism, *so long as the intention these actions are done with is not an intention to deceive oneself* (this was all the non-intentionalist was concerned to deny!). Again, this point seems to get missed, even by non-intentionalists themselves. Thus the well-known non-intentionalist, Alfred Mele, calls the view that self-deceptive beliefs are intentionally acquired/retained, *the agency view*, and the view that they are not intentionally acquired/retained, *the anti-agency view* (e.g., 1998b, 353-354). But the latter view need not imply that the subjects' actions are not responsible (partly or fully) for the acquisition/retention of the self-deceptive belief, as long as those actions are not done with the intention to acquire/retain that belief. So it is misleading to call this 'the anti-agency view'.

---

[6] It is debatable whether desires motivate actions, given that we so commonly speak of non-psychological things as motivating us, e.g., money, power, or the prospect of gaining them. Alvarez (2010, chap.4), for instance, argues that the things we desire, and not our desiring them, are what motivate us. For convenience however, and with misgivings, I will go along with the former very common way of speaking, and 'motivate' can be taken as a technical term if necessary. By a desire motivating an action, I just mean that it was done (partly) because of and was rationalized by the desire.

The point may be expressed in terms of an important distinction that the remarks of Bermúdez and company hint at:

> *Agentivism*: Self-deceivers end up with their unwarranted belief as a result of their own actions motivated by the desire that *p*.
> *Non-Agentivism*: Self-deceivers end up with their unwarranted belief as a result of cognitive mechanisms caused by the desire that *p*.

This distinction, which should not be conflated with the intentionalism/non-intentionalism distinction, can be better understood by noting two different explanatory roles that desires (as in states of desiring) can have.

Perhaps the most commonly recognized explanatory role which desires have is in 'rationalizing' explanations of actions, especially intentional actions or actions done for reasons. As Davidson (1963) noted, when desires feature in such explanations, they help make rational sense of the action, making it intelligible why it was done, helping us to see what the attraction or appeal was in doing it from the agent's perspective. The most familiar case is where an action is made intelligible by the agent's desire by being believed to be a means to its satisfaction (e.g., Jeremy shakes the apple-tree, because he wants an apple, and believes that by doing so he'll get one). Mele (1998a) calls this 'instrumental rationalization'. Note that desires can rationalize other psychological states/attitudes, and not just actions (see Ibid).

But desires and other affective attitudes can enter into other sorts of explanations, where they are referred to as having caused some effect which they *don't* rationalize, generally a bodily or psychological change of some sort as opposed to an action or other attitude. Examples might include a desire for food causing one to salivate at the sight of it, worry or anxiety causing a rise in blood pressure, or suffering a heart-attack from terror or fright.[7] Here, the explicanda are 'mere happenings' which don't need to be made intelligible to us as do the explicanda of rationalizing explanations. There is no demand to ask, 'What was he after?' or 'What good did he see in doing that?', for they are things that happened to the subject, rather than things he did.

---

[7] It can be tricky distinguishing between effects of a desire or emotion, and things that are partly constitutive of them. Is one's increased heart-rate or the pangs in one's stomach an effect or partly constitutive of one's fear for instance? Having a heart-attack, however, surely isn't part of what it is to experience fear/terror.

I wish to avoid getting bogged down in elucidating this distinction more thoroughly, since this interesting yet daunting task would divert us from our primary concerns. But I hope it's evident that there is a genuine distinction to be noted here. (Incidentally, I suspect that philosophers who speak of desires as 'motivating' actions would reserve this term for cases where desires feature in rationalizing explanations. Thus they would say that Jones' desire for food motivated his efforts to get some, but *not* that it motivated his salivating, rather, it merely caused that.) Furthermore, I suggest that we should think of the desire as playing a rationalizing explanatory role in agentive explanations of self-deception, and a non-rationalizing or 'merely causal' one in non-agentive explanations.

Returning to the intentionalism/non-intentionalism distinction, we can now say that intentionalism is clearly a form of agentivism. Non-intentionlist theories, however, can be non-agentivist *or* agentivist. They are agentivist if they explain the belief as resulting from the subject's intentional actions, actions rationalized by the desire that *p*, though the associated intention should not be as intentionalists describe it. They will be non-agentivist if they explain it as resulting from 'sub-intentional mechanisms' 'triggered' by that desire. (The relationships between these distinctions, regarding their combinability, are outlined in box 1).

Box 1

	Agentivism	Non-Agentivism
Intentionalism	Combinable	Not combinable
Non-intentionalism	Combinable	Combinable

Note that a non-agentivist explanation, as I am defining it here, need *not* imply that intentional actions are in no way involved in the causation of self-deceptive beliefs. Suppose *S* desires that *p*, but also needs to come to an accurate judgment on whether *p* and is determined to do so. He searches for evidence both for and against *p*, motivated by the desire to ascertain the truth. However, suppose that *S*'s desire that *p*, unbeknownst to him, primes memories evidentially conducive to believing that *p*, making them more accessible, a 'merely causal' effect of the

desire.[8] These memories then come to mind more easily, and he ends up believing that *p* despite his best intentions, where his actions, say, would have led to him correctly believing that not-*p* were it not for the operation of this biasing mechanism. Here the explanation is non-agentivist, even though his evidence search was an intentional action (done with the intention of determining whether *p*), and was crucial to him forming the belief. This is because his actions or intentions were not to blame for the biased belief, rather, a sub-personal priming mechanism was to blame. His actions would have led to him having the correct belief were it not for the mechanism.

The question of whether self-deceptive beliefs are generally due to the person's actions or to the operation of sub-personal/sub-intentional mechanisms is important, with major consequences for the issue of to what extent people are responsible for their own self-deception. Many non-intentionalist theories emphasize the mechanistic (e.g., Hales 1994; Lazar 1999; Smith 2014). This, however, is an over-reaction to intentionalism's objectionable features, and I here wish to convey the attractions of agentive non-intentionalism, a frequently overlooked theoretical possibility. First, this approach explicitly endorses the common belief that *there are self-deceptive actions*. (Hellman speaks of our 'intuitions that self-deception is a state determined by our behavior' (1983, 120). Also see (Scott-Kakures 2002, 591) and (Michel and Newen 2010, 742)). The grammar we use when talking about self-deception suggests this (Szabados 1974b, 62), as when we use the active progressive 'She is deceiving herself', which communicates that the subject is *actively involved* in her self-deception, as opposed to the stative 'She is self-deceived', which does not directly suggest this. But it endorses this point while avoiding intentionalism, with its threat of paradox.

Furthermore and relatedly, it accommodates the common intuition (e.g., Fernández 2013; Szabados 1974a) that a person's self-deception is something she can be *held responsible* for. The idea that we end up self-deceived due to our own actions seems most congenial to this assumption (in that we can often be held responsible for the consequences of our actions). Note, however, that the current theory will be agentivist in a qualified sense, since it takes the self-deceptive belief to result from the subject's actions *and omissions*. But this will not undermine its

[8] This as least resembles mechanisms postulated by some psychologists where good and bad moods are said to exercise priming effects on 'mood-congruent' memories. So for instance, being in a good mood is said to increase the accessibility of positive memories, making one view the world more positively (Kunda 1999: 249-250). (However, one could argue that viewing the world positively is partly constitutive of being in a good mood, rather than an effect of it.)

accommodating our intuitions about responsibility, since it's undisputed that people can be blameworthy for their omissions. Alternatively, if self-deception is more the result of the automatic, 'merely causal' influence of desire on cognition, we would perhaps be better thought of as hapless victims of the condition, rather than perpetrators.

However, perhaps the main reason for favoring an agentivist non-intentionalism is that empirical studies of self-deception (i.e., cases displaying features A-C) suggest that the 'mediating processes', as psychologists say, between the desire and belief involve intentional actions.[9] Next, a relevant representative study will be presented and interpreted in this fashion.

3. Mediating Processes between the Desire and Belief

In the studies I have in mind, typically a large group of subjects would be presented with evidence suggesting that a proposition is true. Some of them would very much not want it to be true (call them 'stakeholders'), whereas the others wouldn't care either way ('non-stakeholders'). In one such study by Liberman and Chaiken (1992), a replication of one by Kunda (1987), subjects were told that they were participating in a study on 'nonscientists' understanding of scientific and technical information' (1992, 672). Then they were given an article to read purporting that moderate to heavy caffeine consumption greatly increases the risk of fibrocystic disease, a serious breast ailment associated with cancer. The subject pool consisted of women sorted into two categories: 'low-relevance' subjects, who drank no cups of coffee per day (making them non-stakeholders), and 'high relevance' subjects, who drank 2 to 7 cups per day (stakeholders). After reading the material, they filled in questionnaires about it.

From examining these questionnaires, the experimenters found that, as expected, stakeholders tended to be more skeptical of the proposition compared to non-stakeholders. On average their opinions veered in the direction of what they wanted to be true relative to non-stakeholders' views.

Measures were included on the questionnaires to determine the cause of this bias. The article contained four research reports on whether caffeine is linked to this disease, one anti-link

[9] This makes it sounds as though it's an empirical question whether in self-deception the person's actions are responsible for her self-deception. It could be, however, that this is not right at all, and that if there are no self-deceptive actions then there is no *deceiving of oneself*. On this view a non-agentivist theory of self-deception is ruled out logically. All I will say about this is that were it true, it would be even more to the advantage of the approach being developed here.

report, and three pro-link reports. Methodological flaws or weaknesses were included in each report '[to] allow vigilant subjects the opportunity to be critical' (1992, 672). Subjects were then asked to list any strengths and weaknesses that they found in these reports, and their answers to this could be used to assess how the evidence was being treated.

Generally speaking, there are two ways a self-deceiver could respond defensively to unwelcome evidence: she could make a *flight* response, attempting to avoid or ignore it, or a *fight* response, attempting to refute it. These two possibilities were tested for. First, Liberman and Chaiken checked for a *defensive inattention* hypothesis: that stakeholders ended up with their biased views by avoiding attending to the pro-link sections while attending more to the welcome anti-link sections (a flight response). From answers given to a memory test on the details of the reports however, it was found that this was not what happened, a result consistent with other studies. Instead, they found evidence for what they dubbed *biased systematic processing*. Stakeholders attended to both kinds of report equally, but were discovered to have rated the anti-link report as superior. Crucially, whereas non-stakeholders found an equal amount of weaknesses in pro- and anti-link reports, stakeholders listed 'significantly more' weaknesses in the pro-link reports and found *less* weaknesses in the anti-link report, compared to non-stakeholders (1992, 674). As the authors conclude, '[c]ompared with low-relevance subjects, high-relevance subjects were less critical of those parts of the message that were reassuring and more critical of those parts that were threatening' (1992, 675). They exhibited a fight response to the unwelcome evidence, attempting to refute/discredit it.

This study suggests that it was a combination of the subjects' *actions* (being hypercritical of/looking for flaws with threatening evidence) and *omissions* (failing to be critical of reassuring data) that led to the formation of the biased beliefs. In other words, their beliefs resulted from their acts of searching specifically and one-sidedly for such considerations (flaws) that they would like to find. This is consistent with results found in numerous other studies (e.g., Ditto et al. 2003; Ditto & Lopez 1992; Frey and Stahlberg 1986; Holton & Pyszczynski 1989; Lundgren and Prislin 1998; Pyszczynski et al. 1985; Wyer and Frey 1983). For instance, Frey (1981) found that subjects led to believe that they had performed poorly on an intelligence test showed a preference for reading articles critical of intelligence tests over articles supportive of them. (Note that an evidence search may be physical, as with Frey's subjects looking through these articles, or

it may be mental, involving just trying to think of arguments or reasons in support of a proposition.)

Now I don't deny that other non-intentionalists, such as Mele and Scott-Kakures (2002; 2000), would agree that this goes on in self-deception, and they may even have mentioned similar things in their work. However, they have not supposed, or explored the idea, that this might be evidence for a distinctive kind of intention present in self-deception. Identifying this intention promises us a deeper explanation of the phenomenon (as we'll see better in section 5).

The argument that the culpable acts evident in these studies are intentional is a simple and, I hope, a compelling one. First, it is a general truth that acts of searching are goal-directed and intentional. They are done with the intention, or with the hope, of *finding whatever one is searching for*. Thus, if you are searching through your pockets for your key, then you intend to find your key, or if a detective is searching for a missing person, he acts with the intention (or hope[10]) of finding that person. Now the stakeholders in the study seemed to be engaged in a search. But what were they searching for exactly? Apparently, for flaws in the threatening evidence[11] (they found many flaws in the pro-link reports but detected few in the anti-link reports, though they were there to be found). Are we not compelled, then, to attribute to them something like the *intention to find any flaws in the threatening evidence*? Note, however, that although the action component of our explanation is intentional, the omission component need not be, a point that will soon prove useful for rebutting accusations of paradox.

In the next section, this idea will be developed further, and we will see how the desire that *p* can give rise to such intentions. The *rationalizing* role that this desire has in explaining these intentions will also be clarified, which will help make the phenomenon more intelligible to us, and we will address the question of whether there is any danger of paradox with this account.

4. The Associated Intention

The above study and others like it suggest that self-deceptive beliefs result, at least in part, from one-sidedly seeking evidence/considerations that would support one's favored view, while failing to submit such reassuring considerations to critical scrutiny. Note, however, that an intentionalist

---

[10] It is more natural to say 'hope' instead of 'intend' if there is doubt over whether the thing can be found.

[11] I use 'evidence' in a wide sense, which includes abstract reasons and considerations relevant, or seemingly relevant, to an issue.

could accept all this and subsume these findings into her own theory. Accordingly, she could say that self-deceivers search one-sidedly for welcome, *p*-supporting evidence with the intention of *making themselves believe* that *p*, a belief they initially know is false/unwarranted. They want to believe that *p* (to avoid the anxiety of believing/suspecting that not-*p*), think that by searching for evidence/considerations in this biased way they may achieve that, and then act like so to achieve it.

This interpretation of the self-deceiver's behavior would be thought by many to invite paradox. Some would think that it implies that the subjects would at one point concurrently believe that *p* and that not-*p*, though rarely does one see an argument for why the beliefs couldn't be held consecutively instead[12]. The more serious threat of paradox comes from the idea that they intentionally and knowingly search for evidence in a biased way, for how could they be persuaded by a body of considerations that they know is biased? Now perhaps the intentionalist has a reply to this, but I won't pursue this debate. Instead, an indirect argument against the intentionalist will be provided, by showing that an alternative understanding of the self-deceiver's behavior is available which is preferable because it is simpler, psychologically more plausible, and paradox-free. The rationalizing role of the desire that *p* in this account will also be outlined.

On this alternative non-intentionalist theory being recommend, all we need to say to account for the subject's biased behavior, call him Jones, is the following. After the unwelcome evidence indicating that not-*p* is encountered, Jones finds that a previous assumption of his is threatened with being false (though he does *not* form the belief that not-*p* on its basis). Because of his strong desire that *p*, this prospect is a source of distress for him. Because of this, Jones has a heightened interest in any evidence/reasons there might be that would invalidate that unwelcome evidence and remove the threat, for the reason that *it would put his mind at ease to have p confirmed*. He consequently becomes anxious to find such considerations, which, again, if found would allow him to *rest assured* that *p*. Anxiously desiring to find reasons to discount that evidence, he acts with the intention of satisfying this desire, and searches for such considerations. Thus he acts merely with the intention *of finding any weaknesses with the threatening evidence*. Or perhaps he acts with the broader intention *of finding any considerations supportive of p*

[12] It is no more necessary that people who intentionally deceive themselves must at one point believe *p* and not-*p* then it is necessary that people who intentionally kill themselves must at one point be both alive and dead.

(which would include flaws with the evidence against *p*, but could also include considerations independently supportive of *p*). Thus he is searching for considerations he would *welcome* or *like to find*, though his searching need not be intentional under that description.[13] Moreover, in his eagerness to find *p*-supporting considerations, he neglects to submit those considerations to critical scrutiny.

On this account, the culpable behavior (one-sidedly searching for welcome considerations) is explained by an intention (to find any problems with the threatening evidence, or to find any *p*-supporting considerations), and that intention is explained and rationalized by the desire that *p* along the above lines. Again, it should stir no controversy to say that searching for flaws with the threatening evidence, or for *p*-supporting considerations, implies an intention to find such things, just as searching for, say, a missing sock implies having an aim or intention to find that sock. Thus intentional action is a key component of our account.

Now it may be observed that with the intentionalist interpretation, Jones acts with this intention also. On that view, Jones intends to find *p*-supporting considerations too, but the difference is that he is supposedly doing this in the attempt to satisfy his further intention to believe that *p*. What the current view is saying is that the intention to find any *p*-supporting evidence need not be seen as a sub-intention in this larger intentional project for Jones' behavior to be intelligible to us. For just in light of the fact that he desires so much that *p* we can understand why he so eagerly seeks such considerations, simply because his finding them would assure him that what he wants to be true is true, thus putting his mind at ease. This account is therefore simpler than the intentionalist's.

Importantly, this allows us to see how the desire that *p* rationalizes the self-deceptive actions, how it 'makes sense' of them, showing that the desire is not playing a 'merely causal' or mechanistic explanatory role. It is fully intelligible to us why Jones doggedly seeks *p*-supporting evidence, since given that he *desires* that *p*, it would *please* him to find it; it would be *to his relief* (a case of non-instrumental rationalization; see (Mele 1998a)). Furthermore, 'cold biasing mechanisms' do not come into this picture. The biased evidence search is not a case of the confirmation bias being triggered by the desire that *p*. Rather, these subjects are searching for *p*-

[13] Likewise, someone searching for coins under the couch cushions might be searching for something he wants, but that doesn't mean his searching is intentional under that description. That is, he just has the intention to find some coins, not the intention to find something he wants.

supporting evidence *because they want to find such evidence*, to assure themselves that *p* is true. The activity is both initiated *and sustained* by this desire (thus satisfying a desideratum of Scott-Kakures (2000)).

There should be no mistaking how different this account of the self-deceiver's intention is from intentionalism's. Bermúdez (2000, 310) has described various options for how intentionalists can conceive of the self-deceiver's intention, but they all involve the self-deceiver intending to bring it about that he acquires a belief. On this view, the self-deceiver's concern is directed *inward* (figuratively speaking): he is concerned with his own doxastic state. We flatly deny that the self-deceiver has any such intention, and shift towards seeing the self-deceiver as having an *outwardly directed* concern. By this we mean that he is concerned with the truth-value of *p*, not with his own doxastic state. For he is, to reiterate, anxious to put his mind at ease that it's *actually true* that *p*, and so acts in the hope, and with the intention, of finding any considerations that would confirm this, not in the hope of making himself have a belief. This is sufficient to explain the biased behavior visible in the empirical studies, and because he does not have an intention to make himself have a belief, our account avoids any danger of paradox associated with that idea.

The idea that the self-deceiver has an outwardly directed concern makes this account, in my view, more psychologically plausible than the intentionalist's. For us, the self-deceiver's primary motivation is to find evidence that would confirm *p*. For the intentionalist, it is to avoid negative affect. But the issue of whether *p* is clearly the more urgent and significant one. The issue of whether one is likely to get fibrocystic disease, or of whether one's partner is just after one's money, is weightier than the issue of what negative affect one may feel if one believes that these things are true. It is more plausible that our anxieties about the former are what generate self-deceptive biases. We will return to this matter in the final section.

It remains to be shown that this account of the self-deceiver's intention is not blighted with paradox. Intentionalism invites paradox because the self-deceiver is conceived to intentionally act in a deceptive, epistemically suspect manner. Crucially however, on our account, the intentional actions of the self-deceiver are *not* epistemically suspect. Here the point that the bias consists in the self-deceiver's actions *and omissions* becomes important. Recall that the biased behavior of the self-deceiver consists in the following:

1) Seeking welcome considerations
2) Not seeking unwelcome considerations (which includes failing to critically scrutinizing any welcome considerations which are found)

This distinguishes self-deceivers from rational, judicious inquirers, who look for the pros and cons, and not just for pros (or cons) only. Now (1) is an intentional action, since acts of searching are generally intentional. So we may suppose that the subject knows what he's doing here, that is, he knows that he's looking for problems with the threatening evidence, or considerations which may favor *p*.[14] But there is nothing intrinsically suspect about looking for problems with certain evidence, or looking for consideration which might support a certain proposition. Such behavior does not constitute being biased. The bias consists in doing that while omitting to critically assess the welcome considerations that one finds or thinks of, and omitting to put an equivalent effort into looking for considerations supportive of not-*p*. But these omissions need not be intentional. They are instances of neglecting rather than of avoiding or refusing to do something. People naturally put more energy into looking for things they want to find, compared with things they do not want to find. The self-deceiver is just so driven to find *p*-supporting considerations that he neglects to seek and check for not-*p*-supporting considerations, or at least neglects to do so with the same energy. Because these omissions are not intentional, there is no pressure to say he must have been aware of his omissions, and so no pressure to say he was aware he was being biased.

Suppose, however, that he did realize that his evaluation of the matter was one-sided. Would that undermine his self-deceptive belief? It might, and it's plausible to think that reflective self-awareness of this sort is self-deception's antidote. But it also might not. For he still does not know that he would have found problems with the welcome considerations if he had looked for them. Though he might acknowledge his bias, he may still cling to his self-deceptive belief for lack of having any *specific* reasons to surrender it.

A worry for this account, however, relates to whether it is faithful to our own experience of when we do such things. When we reflect on occasions when we acted like so, does it not seem that our searching for welcome considerations proceeds sometimes in a more automatic than deliberate manner? Don't we just immediately seize upon such welcome considerations,

[14] Note, however, that this knowledge typically will not be reflective. The subject typically just does it without thinking about the fact that he's doing it.

without much reflective awareness that we are doing so? It can be admitted that there may be a spectrum of cases here, but the point is one that is true of many sorts of actions. That is, for many actions which are sometimes done in a deliberate or intentional manner, they can at other times be done in a more automatic, spontaneous, or habitual manner (and sometimes in a semi-intentional manner, since these things come in degrees). TV 'channel surfing', for instance, may be done in a very deliberate way, or in a more automatic, 'mindless' way, though in both cases the action is explained by the same desire to find something good to watch. It shouldn't be surprising if actions of searching for flaws with unwelcome evidence are no different in this respect.

5. The Relation Between this and Mele's Processes of Self-Deception

Before finishing, this account should be related to Mele's, who has written a lot on how ordinary or 'garden-variety straight' self-deception (the phenomenon targeted here) comes about. Through this we will see the explanatory depth of the present account, which primarily consists in its unifying power.

Mele mentions a number of apparently *independent* processes that may help explain ordinary self-deception. However, it seems that many of these processes can be subsumed into and unified by the present account; they can be seen as manifestations of the one phenomenon. Consider two such processes. He says that self-deceivers may engage in 'negative misinterpretation', which is when '[o]ur desiring that *p* may lead us to misinterpret as not counting (or not counting strongly) against *p* data that we would easily recognize to count (or count strongly) against *p* in the desire's absence' (2001, 26). The example is given of Don who had his paper rejected from a journal, and who forms the belief, unwarrantedly, that it was wrongly rejected because the reviewers misunderstood a certain crucial but complex point. Or they may engage in 'positive misinterpretation', which is when '[o]ur desiring that *p* may lead us to interpret as *supporting p* data that we would easily recognize to count against *p* in the desire's absence' (2001, 26). He illustrates with the case of Sid who interprets the rebuffs of a woman he's infatuated with as a sign of her playing hard to get.

It seems, however, that both Don and Sid's behavior can easily be understood as being instances of striving to find *p*-confirming/supporting considerations (so that they might rest assured that *p* is true). Don looks for and finds a reason to think that the reviewer got it wrong (as

he hopes), a reason he doesn't scrutinize, while Sid looks for and finds a reason to think that some ostensibly negative evidence isn't negative at all.

Mele also mentions 'selective focusing/attending' which is when '[o]ur desiring that *p* may lead us both to fail to focus attention on evidence that counts against *p* and to focus instead on evidence suggestive of *p*' (2001, 26). This is also easily explicable by us. The reason why the self-deceiver fails to notice evidence for not-*p* is that *he's not looking for it*. He *is* looking for evidence for *p*, however, which is why he notices/focuses on that. How Mele's 'selective evidence-gathering' (2001, 27) fits into our account is even more obvious. If one is looking for evidence for *p* but not for not-*p*, one will end up selectively gathering evidence for *p*. Furthermore, Mele suggests that data consistent with a desired hypothesis can be rendered more 'vivid' or 'salient' by that desire (2001, 29). But again, this 'vividness' can be understood as a consequence of the fact that the subject is searching for that data (think of searching through a cluttered drawer for something. The item you are looking for jumps into your attention and you barely notice the other items).

So many of Mele's self-deceptive processes can be explained in terms of, or understood as manifestations of, the attempt to find any welcome evidence, and moreover, by providing this common explanation, we significantly improve on his account by bringing unity to this otherwise disparate-looking set of factors. We are not denying Mele's explanations but are deepening our understanding of them, by explaining those explanations. Note also that in doing this, we are assimilating them into an *agentive* account of ordinary self-deception. We understand these factors not as mechanisms or the effects of mechanisms, but as actions or the effects of actions. Mele, on the other hand, promoted an incorrect understanding of these processes by calling his view an 'anti-agency' view (2001, 25, 49).[15]

It is doubtful, however, that all the explanatory processes that Mele mentions can be subsumed into this account. Mele's 'FTL model' in particular (an acronym referring to the work of psychologists J. Friedrich, Y. Trope, and A. Liberman), which he offers as a possible, partial explanation of garden-variety self-deception, is one we should perhaps take exception to. As this view has become quite popular with non-intentionalists, championed especially by Scott-Kakures

[15] We cannot definitively say that Mele is a non-agentivist however, since he sometimes describes cases where the subject's intentional actions are clearly responsible for her self-deceptive belief, in particular his Beth case (2001, 18). The problem is just that he gives conflicting signals concerning whether his account is agentivist or non-agentivist.

(2002; 2000), it deserves a sustained discussion. Next we will see that the FTL phenomenon is distinct from the psychological phenomenon we have been looking at here, and reasons will be given for being skeptical about emphasizing it in the explanation of self-deception.

First, a brief exposition (see (Mele 2001) for more detail). The FTL model uses the notion of a 'confidence threshold', which refers to the quantity and quality of evidence required before one accepts a proposition as true. It proposes that confidence thresholds vary depending on what the expected costs are of falsely believing the relevant proposition, costs which the subject desires to avoid. The less costs associated with falsely believing that *p*, the lower one's confidence threshold regarding it (i.e., it will take evidence of a relatively lesser quality/quantity to make one accept it), and the more costs associated with falsely believing that *p*, the higher one's confidence threshold (i.e., it will take evidence of a relatively higher quality/quantity before one accepts it).

Regarding some hypothesis *p* that a subject attempts to evaluate, the expected costs associated with wrongly settling the question in favor of *p* or not-*p* may be 'asymmetric' in magnitude. The subject will then be most concerned with avoiding the costlier error, and thus will have different thresholds for coming to believe *p* than not-*p*. If believing that *p* falsely would be most costly, then his threshold for accepting that *p* will be higher, and it will take relatively strong evidence to convince him of that. Consequently, this subject may be biased towards concluding that not-*p*. He will be disposed to settle on that conclusion more easily. Furthermore, he will be disposed to scrutinize carefully evidence suggesting that *p* for errors, since he is averse to mistakenly concluding that *p*.

To illustrate, consider Frank, a lead engineer overseeing the construction of an aeroplane. Frank has to evaluate whether the plane is flight-worthy. He knows that there are high costs associated with falsely concluding that it's flight-worthy (death, litigation, unemployment). He knows that the costs of falsely believing that it's not flight-worthy are much less significant (resources needlessly wasted). He desires most to avoid the costlier error. So his confidence threshold for accepting that the plane is flight-worthy is higher. Thus he is disposed to conclude that it's not flight-worthy more easily, and he will scrutinize carefully any evidence (e.g., reports) suggesting that it is flight-worthy.

The first point to make about such cases is that it is not clear that the thresholds in question here are belief thresholds as opposed to *acceptance* thresholds in the distinct sense

explained by Michael Bratman (1992) and others. For acceptance is supposed to be influenced by precisely the sorts of practical considerations described in the FTL model. And if they are acceptance thresholds then this might undermine the relevance of the FTL phenomenon to self-deception insofar as self-deception involves belief. So just as, say, if one were driving on a remote and dangerous road one might 'accept', 'presume' or 'work on the assumption' that one is going to meet another car on the blind turns while possibly believing that it's quite unlikely to happen, the engineer might just be accepting that that the airplane is not yet flight-worthy without that necessarily meaning that he believes it's not yet flight-worthy.

But let us grant, for the sake of argument, that the thresholds are belief thresholds. This example then shows that 'FTL motivation' (motivation to avoid doxastic error costs) does not in itself lead to self-deception. For even if Frank ended up with a belief that the plane is not yet flight-worthy when it is, though he might be guilty of over-cautiousness, he would hardly be guilty of self-deception. Certainly this would not be a case of 'garden-variety straight' self-deception since this belief may not be one which he desires to be true (it could make little difference to him when the plane becomes flight-worthy). To see how the FTL phenomenon is relevant to self-deception, we must look at how desires relating to the truth of the relevant proposition can give rise to doxastic error costs.

Mele says that where someone has a strong desire regarding the truth of a proposition, this will affect what costs there are for him in falsely rejecting/accepting it. Most importantly, where *S* desires that *p*, if *S* falsely believes that not-*p* he will incur *negative affect* (2001, 35-36), but not if he falsely believes that *p*. *S*'s desire to avoid this cost will then give him a relatively low threshold for believing that *p*, and will make him vigilant with regard to weaknesses in evidence supporting not-*p*.

Note, however, that the thing which *S* supposedly wants to avoid here, negative affect, would be incurred whether *S* believes that not-*p falsely or truly*. Experiencing this affect is not dependent on the belief being a *mistaken* belief. Mele acknowledges this point (2001, 35), but it implies that this does not exemplify FTL motivation: it is not a matter of being motivated to avoid the costs of making some sort of *error*. When we realize this, we see that this is just an old idea dressed up in new garb: the idea that in self-deception, the subject is motivated to avoid the anxiety associated with believing something, which biases his reasoning.

This idea that the self-deceiver's concern is an 'inwardly focused' one of avoiding anxiety is one this paper has been trying to move us away from; it conveys an implausible picture of the self-deceiver's primary worry. For when one, say, encounters evidence suggesting that one may be seriously ill, one does not, first of all, become worried about the anxiety one will suffer in believing this. One's worry primarily concerns *being seriously ill*, with all the implications that would have for one's future (or lack of one). Surely any relatively trifling worries about what negative affect one might suffer in believing this (falsely or truly) will be secondary to this, if they arise at all (Bach 1981, 353; Szabados 1974c, 470). And it should come as no surprise if this outwardly directed concern has the power to distort one's thinking, independently of any inwardly directed concerns about avoiding emotional distress.

It is time that we questioned the dominant yet implausibly solipsistic idea that the self-deceiver's primary concern is with avoiding anxiety rather than with the truth/falsity of the proposition he's self-deceived about. For we know only too well that our having a major stake in some proposition being true leads naturally (when the proposition is thrown into doubt) to an intense desire for there to be evidence that it is true, which leads us to search doggedly for such welcome evidence, so that we might be relieved that it is true, and this motivational complex by itself can generate a biased belief. It is this motivational complex that is the main engine of ordinary self-deception, and a person so concerned cannot be rightly described as being concerned about negative affect. She is concerned about whether something is true or not.[16]

The cost of negative affect is what Mele emphasizes the most, but other sorts of costs have been claimed to be generally associated with falsely rejecting desired propositions. It has been claimed, for instance, that this may often involve missing important opportunities (see Scott-Kakures 2002, 594). Consider a subject, *S*, in Frey's experiment, who convinces himself that intelligence tests are inaccurate measures of intelligence after having performed poorly on one. If *S* falsely believed that intelligence tests are accurate, he could take himself to be less intelligent than he is. He might then assume that he is unfit for pursuits (e.g., a certain career) he actually is fit for, and could thus miss opportunities. Granted. But note that there are potential costs, no less serious, with the converse error. If he falsely believed that intelligence tests were

[16] This is not to deny that self-deception results in a reduction of negative affect. But we deny that the self-deceiver's purpose, in behaving as he does, is to secure such a reduction. His purpose is to find *p*-confirming evidence, and his concern is with the truth-value of *p*. The reduction of negative affect is a by-product of his achieving that purpose.

inaccurate, he might assume that he is smarter than he is. This could lead him to take up pursuits that he's unable for, leading to eventual failure, disappointment, and perhaps wasted years. Both overestimating and underestimating your intelligence carries risk. Or consider a subject in Liberman and Chaiken's experiment, who convinces herself that the studies linking caffeine consumption to fibrocystic disease are problematic. Wouldn't there be significant costs with falsely *accepting* that proposition, in that she could be dismissing a genuine danger?

Advocates of the FTL model have thought that it could serve as a general explanation of paradigmatic self-deception, because it has seemed to them that we *generally* regard falsely rejecting desired propositions as more costly than falsely accepting desired propositions. But it's not evident that this is true. Costs can often be identified on both sides. We would thus need evidence that in cases of self-deception generally, subjects associate high (and higher) costs with falsely rejecting the welcome proposition, if there is to be any hope of the FTL theory delivering the goods. Without that, it is little more than a conjecture. The protagonist motivational states in the current explanation however (e.g., the desire to find problems with the threatening evidence, or to find *p*-supporting evidence), are more likely to be generally present in the relevant cases (who doesn't want to find *p*-supporting evidence when a proposition, *p*, that one strongly wants to be true is thrown into doubt?). And furthermore, they are likely to be a more powerful force in the subject's mind than the likes of desires to avoid anxiety or to not miss opportunities, and so are more likely to be the main culprit responsible for motivating the biased behavior.

Acknowledgments

Thanks are due to Maria Baghramian, Martin Davies, Paul Doody, Johannes Roessler, Markus Schlosser, Matt Soteriou, Rowland Stout, and to the anonymous reviewers of this and other journals for comments and criticisms on previous drafts of this paper. Thanks also the audiences for their feedback at a seminar in 2011 at Warwick University, UK, and at the *1st International Krakow Conference in Cognitive Science* in 2012 at Jagiellonian University, Poland, where some of these ideas were presented.

References

Alvarez, Maria. 2010. *Kinds of Reasons: An Essay in the Philosophy of Action*. Oxford: Oxford University Press.

Bach, Kent. 1981. "An Analysis of Self-Deception." *Philosophy and Phenomenological Research* 41(3): 351-370.

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

Bratman, Michael. 1992. Practical Reasoning and Acceptance in a Context. *Mind* 101: 1-15.

Connolly, Emer. 2008. *Lying Eyes and the Hitman for Hire*. Gill and Macmillan.

Davidson, Donald. 1963. "Actions, Reasons, and Causes." *Journal of Philosophy* 60(23): 685-700.

Deweese-Boyd, I. D. 2006, revised 2012. "Self-Deception." *Stanford Encyclopedia of Philosophy*.

Ditto, P. H., and D. F. Lopez. 1992. "Motivated Skepticism: Use of Differential Decision Criteria for Preferred and Nonpreferred Conclusions." *Journal of Personality and Social Psychology* 63(4): 568-584.

Ditto, P. H., G. D. Munro, A. M. Apanovitch, J. A. Scepansky and L. K. Lockhart. 2003. "Spontaneous Skepticism: The Interplay of Motivation and Expectation in Responses to Favorable and Unfavorable Medical Diagnoses." *Personality and Social Psychology Bulletin* 29(9): 1120-1132.

Fernández, Jordi. 2013. "Self-Deception and Self-Knowledge." *Philosophical Studies* 162(2): 379-400.

Frey, Dieter. 1981. "The Effect of Negative Feedback about Oneself and cost of Information on Preferences for Information about the Source of this Feedback." *Journal of Experimental Social Psychology* 17(1): 42-50.

Frey, Dieter, and Dagmar Stahlberg. 1986. "Selection of Information after Receiving More of Less Reliable Self-Threatening Information." *Personality and Social Psychology Bulletin* 12: 434-441.

Hales, S. D. 1994. "Self-Deception and Belief Attribution." *Synthese* 101(2): 273-289.

Hellman, Nathan. 1983. "Bach on Self-Deception." *Philosophy and Phenomenological Research* 44(1): 113-120.

Holton, Brynda, and Tom Pyszczynski. 1989. “Biased Information Search in the Interpersonal Domain.” *Personality and Social Psychology Bulletin* 15(1): 42-51.

Kunda, Ziva. 1999. *Social Cognition: Making Sense of People*. Cambridge, Mass.: MIT Press.

Kunda, Ziva. 1987. “Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories.” *Journal of Personality and Social Psychology* 53(4): 636-647.

Lazar, Ariela. 1999. “Deceiving Oneself of Self-Deceived? On the Formation of Beliefs “Under the Influence’.” *Mind* 108: 265-290.

Liberman, Akiva, and Shelly Chaiken. 1992. “Defensive Processing of Personally Relevant Health Messages.” *Personality and Social Psychology Bulletin* 18(6): 669-679.

Lundgren, S. R., and Radmila Prislin. 1998. “Motivated Cognitive Processing and Attitude Change.” *Personality and Social Psychology Bulletin* 24(7): 715-726.

Lynch, Kevin. 2016. “Willful Ignorance and Self-Deception.” *Philosophical Studies* 173(2): 505-523.

Lynch, Kevin. 2012. “On the ‘Tension’ Inherent in Self-Deception.” *Philosophical Psychology* 25(3): 433-450.

Mele, A. R. 2001. *Self-Deception Unmasked*. Princeton; Oxford: Princeton University Press.

Mele, A. R. 1998a. “Noninstrumental Rationalizing.” *Pacific Philosophical Quarterly* 79(3): 236-250.

Mele, A. R. 1998b. “Motivated Belief and Agency.” *Philosophical Psychology* 11(3): 353-369.

Michel, Christoph, and Albert Newen. 2010. “Self-Deception as Pseudo-Rational Regulation of Belief.” *Consciousness and Cognition* 19: 731-744.

Pyszczynski, Tom, Jeff Greenberg and Kathleen Holt. 1985. “Maintaining Consistency between Self-Serving Beliefs and Available Data: A Bias in Information Evaluation.” *Personality and Social Psychology Bulletin* 11(2): 179-190.

Smith, D. L. 2014. “Self-Deception: A Teleofunctional Approach.” *Philosophia* 42(1): 181-199.

Scott-Kakures, Dion. 2002. “At ‘Permanent Risk’: Reasoning and Self-Knowledge in Self-Deception.” *Philosophy and Phenomenological Research* 65(3): 577-603.

Scott-Kakures, Dion. 2000. “Motivated Belief: Wishful and Unwelcome.” *Noûs* 34(3): 348-375.

Szabados, Béla. 1974a. “The Morality of Self-Deception.” *Dialogue* 13(1): 25-34.

Szabados, Béla. 1974b. “Self-Deception.” *Canadian Journal of Philosophy* 4(1): 51-68.

Szabados, Béla. 1974c. “Rorty on Belief and Self-Deception.” *Inquiry* 17(1): 464-473.

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

Wyer, R. S., and Dieter Frey. 1983. “The Effects of Feedback about Self and Others on the Recall and Judgments of Feedback-Relevant Information.” *Journal of Experimental Social Psychology* 19(6): 540-559.