palgrave communications

HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

DOI: 10.1057/s41599-017-0017-0 OPEN

Self-deception in and out of illness: are some subjects responsible for their delusions?

Quinn Hiroshi Gibson[1]

ABSTRACT This paper raises a slightly uncomfortable question: are some delusional subjects responsible for their delusions? This question is uncomfortable because we typically think that the answer is pretty clearly just 'no'. However, we also accept that *self-deception* is paradigmatically intentional behavior for which the self-deceiver is *prima facie* blameworthy. Thus, if there is overlap between self-deception and delusion, this will put pressure on our initial answer. This paper argues that there is indeed such overlap by offering a novel philosophical account of self-deception. The account offered is independently plausible and avoids the main problems that plague other views. It also yields the result that some delusional subjects are self-deceived. The conclusion is not, however, that those subjects are blameworthy. Rather, a distinction is made between blameworthiness and 'attributability'. States or actions can be significantly attributable to a subject—in the sense that they are expressions of their wills—without it being the case that the subject is blameworthy, if the subject has an appropriate excuse. Understanding delusions within this framework of responsibility and excuses not only illuminates the ways in which the processes of delusional belief formation and maintenance are continuous with 'ordinary' processes of belief formation and maintenance, it also provides a way of understanding the innocence of the delusional subject that does not involve the denial of agency.

[1] Department of Philosophy, University of California, Berkeley, CA, USA. Correspondence and requests for materials should be addressed to Q.H.G. (email: qhgibson@berkeley.edu)

Introduction

In this paper I intend to raise a somewhat uncomfortable question: are at least some delusional subjects *responsible* for their delusions? The question strikes us as uncomfortable at least in part because we think the answer is just pretty clearly 'no'. Nevertheless, I will argue that at least some delusional subjects are responsible for their delusions. My argument will be as follows: When we consider the dynamics of a related phenomenon—to wit, self-deception—we will see that there is enough overlap between them to ground the judgment that self-deception is implicated in the formation and maintenance of at least some delusions. In order to show this, I will first offer my own account of self-deception. We typically think self-deceivers are responsible, and my account captures the sense in which this is correct. I then argue that, according to my account, at least some delusional subjects are self-deceived. Importantly, I believe that this can be shown to be the case without leading us to the judgment that delusional subjects are blameworthy for their delusions. In order to thread this line, I will appeal to the distinction between what I will call 'attributability' (roughly[1] following Shoemaker (2011) and Watson (2004b)) and blameworthiness. I will argue that while self-deceivers are typically responsible both in the sense that their self-deception is attributable to them and in the sense that they are blameworthy, delusional subjects, even when they are self-deceived, are typically only responsible in the sense that their delusions are attributable to them. Why this should be so will be made clear by consideration of the details of my own view of self-deception, as well as the details of the delusions which I consider.

A little bit more about the significance of our question: lying behind the seemingly ordinary idea that delusional subjects are not responsible is the idea that delusions are somehow beyond the scope of ordinary interpersonal understanding. Karl Jaspers (Jaspers, 2007, pp 174–175) famously distinguished this kind of understanding from what he called 'explanation'. Jaspers was aware that psychiatry was partly a natural science, and partly a human science. Explanation is what natural science does: it uses objective empirical methods to elucidate causal structures. Understanding, on the other hand, is unique to human science, and uses ordinary interpersonal imagination and other 'subjective' methods to appreciate the experiences of subjects 'from the inside' (Kendler and Campbell, 2014, p 1). Jaspers nevertheless found that understanding could sometimes break down in the face of more extreme symptoms. The prevailing ethic in contemporary medical psychology seems to agree with him. The idea is to regard patients suffering extreme symptoms as deserving of compassionate treatment, but also as nevertheless, at some ultimate level, perhaps *beyond understanding*—or as Jaspers himself put it, 'un-understandable'. I will not (and cannot) argue that understanding does not break down in the face of some extreme conditions, but it is my view that we should push the boundaries of such understanding as far as they can go in the hopes of coming to grips with how best to understand, in ordinary humanistic terms, what is going on in certain forms of mental illness. I will return to some of the consequences of my argument for the understandability of delusions below under 'Responsibility and delusion'.

So, the uncomfortable nature of our question belies a commitment to extending ordinary human understanding—and indeed, the boundaries of the moral community—as inclusively as we can. For non-experts, our understanding of delusions depends on a highly elaborated medical practice to which we are largely outsiders. And I wish to take seriously the critical idea that practices such as institutionalized medicine and the knowledge which they enable often conceal dynamics of unequal power (Foucault, 1969).[2] This behooves us to be sensitive to the tacitly normative aspects of the explanatory categories appealed to by such practices (categories such as *delusional*), and the effects that such categorization may have on those who are subject to it. Whether someone who is suffering from delusions is—and whether it is appropriate eo ipso that they should be made to *feel like*—a non-agent, a passive sufferer, or someone who is generally non-responsible, are philosophical questions, and answers to them should not be implicitly imported along with the very idea of a delusion. This discussion is an attempt to provide a philosophically sound way of broaching these questions, and to temper the temptation to give too-easy answers to them.

So, I think our question is important. As I said, to go about answering it, I will argue that there is overlap between delusion and self-deception. More precisely, I will argue self-deception can play a role in the formation and maintenance of delusions. But as I said (and I hope to be able to illuminate why this ought to be so) we typically judge self-deceivers responsible for their self-deception. We are also, as I said, pulled towards the claim that delusional subjects are not responsible for their delusions. Taken together with the thesis that I want to argue for, this suggests the following triad:

1. Delusional subjects are *not* responsible for their delusions
2. Self-deceived subjects *are* responsible for being self-deceived
3. There is overlap between self-deception and delusion

If (1) and (2) are read as generics (and they certainly should not be read as universal generalizations), then the triad is not, strictly-speaking, inconsistent. But it points to the need to say something about how we should think of the identified cases of overlap with respect to responsibility. Are they, in this respect, more self-deception-like or more like typical delusions? I hope to be able to make clear, by way of appeal to my account of self-deception, why we should go for the former and not the latter. With that in mind, let us turn to my account of self-deception.

Self-deception as omission

Amongst philosophers there is still a lively debate going on concerning the correct analysis of self-deception, and I will not be able to settle that debate here. But I do have a view to offer (which I develop more extensively elsewhere). To begin, let's consider an example:[3]

> **A:** A is an academic who is self-deceived about the quality of his own work. A is unhesitant about advertising what he takes to be his own brilliance to others, but it is clear to his colleagues and everyone familiar with his work that the work is flimsy. Nonetheless, A badgers OUP to put out a volume of his collected papers. He avoids situations where he might have to confront his work's obvious shortcomings, and when he does encounter criticism he dismisses it as jealous or vindictive. It is clear that A longs deeply for the respect and admiration of his colleagues, but it is equally clear that his pursuit of it is self-undermining.

What is the right way to describe what is going on with A? A natural place to begin, and the view which most philosophical debate about self-deception takes as a starting point, is to think of self-deception as the intrapersonal analogue of ordinary other-deception. I've called this view 'the naïve view':

The naïve view of self-deception: A is self-deceived that p just in case A believes that not-p and A has acted intentionally so as to cause A to believe p

According to the naïve view, A believes that his work is flimsy, and he has somehow managed to act so as to cause himself, on that basis, to come to believe that his work is not flimsy. There is

a way in which this captures the phenomenon. A seems to believe both things, and the more comforting belief seems to be a defensive response to the more sobering one. But there's a big problem. How could someone ever manage to do what the naïve view describes successfully? *How* is an agent to act intentionally so as to get herself to have a belief when she also already believes that very belief to be false? Even moderate doxastic voluntarists would admit that an agent cannot come to believe anything at all just at will, and it seems if anything constrains what one is able to believe at will, it is precisely what else one knowingly believes. Even if there is no other obstacle to my coming to believe that it is raining, the fact that I already believe, and am going to continue to believe, that it is not raining, and that I know this about myself, seems more than enough to prevent me from believing that it is raining. There seems to be three moving parts to the problem: belief, intentional action, and psychological unity. *If* the self-deceived agent could somehow pull off an intentional act of getting himself to believe what he also believes to be false, would this not undermine his psychological unity? This seems like a process which the agent's psychological unity would suffice to prevent. *If* what the self-deceived sufficiently unified agent manages to bring about in himself is a genuine *belief*, would he not have to have done it non-intentionally? *Mutatis mutandis*, if the sufficiently unified self-deceived agent *really intends* to deceive himself, how can the deception involve the bringing about of a genuine belief?

Philosophers have dealt with this so-called 'dynamical' problem of self-deception in different ways. Some have downgraded the self-deceptive act from fully intentional to something less than that (Mele, 1997); others have downgraded the self-deceptive state from fully doxastic to something than that (D'Cruz, In prep.; Gendler, 2007; Darwall, 1988); still others have thought it would suffice to give up on a certain degree of psychological unity (Davidson, 2004; Pears, 1984).

All of these approaches have their advantages and disadvantages. For most views on offer, the problem is quite simply that they do not adequately address the dynamical problem. The scope of the present inquiry forbids going into all the details, but this is worth illustrating, so let's consider one such view.

One particularly popular strategy has it that self-deception is a kind of *pretense*. I will focus on Stephen Darwall's (1988) version of the view, but Tamar Gendler (2007) has also proposed a similar view, and Jason D'Cruz (In prep.) has a revision of Gendler's view. According to the self-deception-as-pretense view, when one is self-deceived about p one need not believe it (although one does typically believe its negation). Rather, one is engaged in an elaborate *pretense* according to which p is the case. One acts *as if* p were true. But unlike in ordinary pretense, where one also believes that one is engaged in pretense, when one is self-deceived, one is also engaged in a *second-order* pretense about one's first-order pretense: one behaves *as if* one is not merely behaving as if p.

Darwall claims that the self-deceived agent need not literally believe the thing that he is supposedly self-deceived about. Perhaps he just *thinks* various thoughts that amount to a kind of elaborate pretense to the effect that the thing in question is the case. But, in ordinary cases of pretense, we *know* that we are pretending. So, in order for the pretense to have the desired psychological results (preservation of self-image, successfully avoiding facing up to painful realizations, etc.) it seems that the nature of the pretense itself has to be concealed. As Darwall puts it, '[This is] not simply the first-order pretense involved in fantasy, but also the second-order pretense that…pretensions are real. When the self-deceiver plays the role of fool to himself, he must also pretend that he is not playing that role' (1988, pp 414–415).

I take it that the reason that the second-order pretense is necessary is to conceal from the self-deceiver the fact that he is engaged in pretense. But now we have to face squarely the question of how someone could ever manage to get himself into *that* state in the first place. And further, the purpose of engaging in the pretense must not be simply to sharpen theatrical skills or for merry diversion. Plausibly the purpose is something self-directed and psychological such as, again, the preservation of self-image, or to avoid facing up to some painful facts. But this is the sort of thing that just cannot be achieved by pretense if one knows that one is pretending. So the purpose of engaging in the pretense, whatever it is precisely, will often be something which cannot be achieved unless it is hidden from the agent.

But if the reason for engaging in the second-order pretense is to make the first-order pretense more credible—i.e., to conceal it as pretense—how are we to make sense of the act of engaging in that second-order pretense without attributing to the agent the very knowledge that would undermine its aim? It seems that the agent must intend to get himself into the state of engaging in both pretenses for the sake of achieving a psychological end, but he must somehow manage to do this without revealing to himself that this is what he is doing. If the problem with intentionally trying to acquire a belief that one also believes to be false has to do with the fact that (acquiring this kind of) belief is not under control of the will, the problem here is that what one is able to conceal from oneself about what one is doing is not under control of the will either. For that matter, why would second-order pretense do the trick, even if we could pull it off? Would we not need third order pretense, and so on, indefinitely?

This is merely illustrative, but some of the difficulties there are generalizable to other views and I take the difficulties to be serious enough to warrant a different approach. According to my own view, there is no single act which is *the act of self-deception*. This is because, on my view, it is not always crucial to self-deception how the self-deceptive belief comes about. Often what is crucial is how that belief is maintained. My view says:

Self-deception as omission: An agent is self-deceived that p if she believes p and intentionally omits to seek, recognize, or appreciate externally available evidence for not-p, for reasons which ultimately relate to her desire that p be true, in a way which enables the maintenance of the belief that p.[4]

By severing the connection between some distinctively self-deceptive process of belief formation, and the resulting self-deceptive state, we can avoid the dynamical problem. My view also captures—as some others do not—the sense in which self-deceivers are responsible: it describes a distinctive kind of motivated epistemic failure which (as we will see shortly) can be grounds for at least a couple of varieties of responsibility judgments.

My view also captures what is going on with A. From the vignette we have seen, we do not know how A came to his belief about the quality of his work, but on my view that does not much matter. A's belief in the quality of his work may have been well-founded at one point. Perhaps he used to be a big fish in a small pond, outperforming other undergraduates at the small state school where he studied, but as he advanced through his career the abilities of those he was surrounded by rose consistently, while his remained stagnant. Or maybe he was never really cut out for academic work and his belief was formed directly and unconsciously as a way of dealing with the stresses of academic life. According to Self-deception as Omission, it does not much matter. What matters is that *now* the belief is manifestly defeated by evidence which is readily available, but A is impervious to that evidence because he *prefers* to continue to believe as he does—that is, he has a desire that p be true—and this motivates him to forswear looking into the matter any further. This is why he

avoids confrontations with others in his field. He omits to do whatever it is precisely that the epistemic norms say that he ought to do in order to bring his belief into proper conformity with the evidence.

What does it mean for the agent's reasons for omitting to seek, recognize, or appreciate evidence to *relate* to her *desire* that *p* be true? I will say that a subject is *emotionally entangled* with a proposition *p* if she is liable to satisfaction or dissatisfaction when *p* is believed to be true or false. To *desire* that *p* be true is for satisfaction to accompany the belief in *p* and dissatisfaction the belief in not-*p* (as opposed to the other way around). Of course, the belief that *p* does not formally satisfy the desire that *p* be true. Rather, the formal object of that desire is the truth of *p* itself. So the satisfaction and dissatisfaction in question for emotional entanglement is not formal, nor is it experiential.[5] It is, we could say, *representational*. It is the kind of satisfaction or dissatisfaction that obtains when the subject's take on the way the world is more or less closely approximates the way the subject thinks the world *ought* to be. This is obviously a matter with motivational efficacy, but it can be so without having a readily identifiable experiential component.

There is admittedly something metaphorical in talk of entanglement, but this expression captures something about the way in which the interaction between, and layering of, desires can produce a complicated web-like structure. There are many ways in which I may desire the truth of some proposition. I might desire it to be true for its own sake (such as I might desire to have a good relationship with someone); I might desire it to be true for the sake of something else (such as I might desire my car to function well); I might desire that something be case *rather than something else* (such as I might desire my partner's infidelity as an explanation for her growing distance over my own emotional unavailability). These are all things I can be self-deceived about because they are things in the truth of which I can manifest an emotional interest.[6, 7]

My view departs from the naïve view in one very crucial way by finding what is distinctive about self-deception, not in the process of belief formation, but in the dynamics of belief maintenance. It is usually taken for granted in the self-deception literature that nothing should count as a self-deceptive state that does not arise from some distinctively self-deceptive process, that there is a constitutive connection between self-deception as a process, and self-deception as the product of that process. Call this thesis the *constitutive connection thesis*. Now, if the constitutive connection thesis is true, it is clear what the object of philosophical analysis ought to be: If what *makes* something self-deception is how it comes about, we had better figure out how it comes about. Of course, there is a trivial reading on which the constitutive connection thesis is true: every particular that comes about as the result of a process is constitutively connected with that process. You do not get goulash by baking pie. However, with self-deception, the connection between process and product is thought to be more intimate than that. The fact that some process was a process of deceiving oneself confers on the resulting state the status of self-deception. Suppose that state is a belief. There are lots of ways to get a belief, many of which are epistemically respectable. Only one way (or some privileged set of ways) of getting that belief is a self-deceptive way. Even though whatever way the goulash comes about will trivially be a goulash-making process, it does not much seem to matter (purist intuitions about goulash with science-fiction pedigree notwithstanding) from the standpoint of assessing whether something *is* goulash which *particular* process resulted in it. This does not seem to be true for self-deception. Or so the thought goes.

My view involves the denial of the constitutive connection thesis. The way I wish to cash this out it is to appeal somewhat more perspicuously to the distinction between belief formation and belief maintenance. We ought not to think that a perfectly clear *temporal* line can be drawn to distinguish between processes of belief formation and processes of belief maintenance: When is the belief formation process over, and when does the process of belief maintenance begin? I will say that a process (or part of a process)[8] is one of belief formation if the belief counterfactually depends on it for the agent's credence in it to increase; and that a process is one of belief maintenance if the belief counterfactually depends on it for the agent's credence in it not to decrease.[9] On this way of thinking about it, some processes will (relative to a given belief) clearly be processes of belief formation (such as, with respect to the belief that it is raining, the perceptual experience of seeing the rain outside my window); others will clearly be processes of belief maintenance (such as, with respect to the belief that it is raining, not encountering any evidence to the contrary in the meantime); and a great deal will be both (such as, looking again and seeing that it is still raining, as opposed to seeing that it is not).

As I suggested could be the case with A, I think there are cases of self-deception where one initially had good evidence for what one believes, but where the evidential situation has since changed, and this change has gone unnoticed for some motivationally biased reasons. To give another example, suppose I believe that I am popular with the kids at school. Maybe I *was* popular with the kids, but kids are fickle, and they have since turned on me. It seems to me that I might be self-deceived if the reason that I continue to believe as I do is because I am impervious to the manifestly available evidence on account of my preference for continuing to believe as I do. We can suppose that once I have reached the point where the kids have turned on me, my credence in the belief that I am popular is not increasing and eo ipso it does not counterfactually depend on me doing or not doing anything in order to increase. But my belief does counterfactually depend on my doing something—or more precisely, not doing something—in order for my credence not to decrease. The evidence is manifestly there, and confronted with such evidence a rational agent would revise her beliefs. What's going on with me? I'm self-deceived! I am intentionally omitting to do what is necessary to bring my beliefs in line with the evidence.

What is the nature of this intentional omission? One might wonder the following:[10] If all the agent is doing is maintaining her belief—especially if it is done via omission—in what way is self-deception an intentional phenomenon? For it to be intentional, would not the agent have to be *knowingly* maintaining her belief against available evidence? Does not Self-deception as Omission face a revised version of the dynamical problem? This worry is actually two distinct worries, and I take them both in turn.

The first worry is that my view would not be able to capture what is intentional about self-deception without facing a version of the dynamical problem. The problem is thought to arise because for something to be self-deception, not only does *something* about it have to be intentional, but it seems that the violation of epistemic norms—the irrationality itself—has to be somehow intentional. This, I take it, is what makes this worry seem like a version of the dynamical problem.

Now, of course, belief maintenance that flies in the face of manifest evidence to the contrary cannot be fully knowing. But it need not be in order to be fully intentional. Here I wish to make a move which Al Mele makes in giving his account (1997) of self-deception. Mele distinguishes, in effect, between intending something *de re* and intending it *de dicto*. According to Mele's view the self-deceiver intends to do something which is an act of deceiving herself without intending to deceive herself as such.[11] This, Mele thinks, recovers what is intentional about self-deception, all the while deflating it to avoid the dynamical problem.[12] I think this is precisely the right move to make *after* we have given

up on the constitutive connection thesis. Once we are talking merely about belief maintenance, and not belief formation, the imagined challenge for my view is not to give a psychologically coherent account of some process, but rather to recover a node of intentional agency which is also recognizably an epistemic failure. The answer to this challenge is pretty straightforward on my view: the agent intentionally (*de re*) omits to seek, recognize, or appreciate externally available evidence for reasons that are motivationally biased, even if she does not intend to do these things as such. So, after it has become clear to everyone else that I am no longer popular, I may simply *do nothing* by way of further investigation into the matter and thus continue to believe as I do. The omission may, of course, not be total. It could be that on occasion I do *encounter* evidence but I omit to *put it together* or to *engage with it as evidence*.[13] So long as my lack of further effective epistemic engagement is motivated by a desire that it be true that I am popular, then I will be guilty of self-deception according to my view. And of course this need not be a one-off affair. My desire that a certain proposition be true may cause me to forego epistemic engagement on many separate occasions.

The second worry here has to do more directly with my appeal to omissions. It's not true in general that every time I omit to do something, I do so intentionally. If someone in the next room requires aid, but I do not know it, then it seems I do not intentionally omit to aid them if I go on reading my book. But this is precisely the sort of knowledge which is denied to self-deceivers on pain of falling into the dynamical problem. My failure to seek, or my failure to engage with, evidence against what I believe cannot be motivated by *knowledge* that it is evidence against what I believe. But the norms that the self-deceiver violates are, in the first instance, epistemic norms, whereas the norm that requires me to render aid is a moral norm. Moral norms may fail to apply to agents who do not have the right knowledge, assuming the agent is not culpable for her ignorance itself—this seems to be a version of ought-implies-can. But epistemic norms cannot be wriggled out of in the same way, especially if they are norms that require an agent to form a particular belief (against a background of evidence and other beliefs). Epistemic norms say how one ought to conform one's beliefs to evidence, or to one's other beliefs, and it is no violation of ought-implies-can that the agent not already have the target belief. If it were, then it would never be epistemically required that anyone form any belief that they do not already hold, no matter how strong the evidence. Ignorance itself cannot be—at least not in the same straightforward way—grounds for claiming that an epistemic norm does not apply as it can be with moral norms. So, while it is plausible that some knowledge is required for an omission to count as a violation of a moral norm, it is not plausible in the same way for omissions which are violations of epistemic norms.

What I want to claim next is that the self-deceiver's motivated epistemic failure is a form of mental agency. We interpret him as having some motivations—wanting to believe well of himself, wanting acclaim in the profession, and so on—and those motivations underlie his failure to bring his belief into conformity with the evidence that is available to him. The state that he thus ends up with is, I will say, *a manifestation of his will*. Allow me to elaborate this by distinguishing two different kinds of responsibility.

What kind of responsibility?

To make the distinction I want to make, let's consider a very simple example. Suppose you step on my foot. Naturally, perhaps, I may want to blame you. First things first. First: are you a *candidate* for moral assessment? Are you a member of the moral community towards whom attitudes like praise and blame are ever appropriately directed? One way to get at this is to ask: Are you a normal adult human being who can recognize and respond to reasons for action? If you are a child, or a paramecium, or—as we too often say—if you are *insane*, you do not have that capacity and we say you are *exempt* (Strawson, 1962, p 3) from assessment altogether. (Obviously there are some forms of insanity which ground exemptions of this kind. I do not think having delusions, on its own, is one such form. What it means to be 'insane' in this sense is, in part, the topic of current discussion.)

Suppose you are the right kind of creature with the right kinds of capacities to be a candidate for moral assessment. Still, that does not settle the question of whether you are blameworthy. We must now ask whether you are *excused*. There are at least two varieties of excuses:[14]

1. *Strong* excuses work by undermining the agent's ownership of the state or action itself. If the action is not *yours* in the right way you are not blameworthy for it. So, if you stepped on my foot because you were shoved by a passerby, you are not blameworthy because, strictly speaking, stepping on my foot was not something that you yourself *did*.
2. *Weak* excuses block the step from an agent's ownership of the state or action to blameworthiness. If you have a blind spot, and my foot happened to be in it while you were trotting on your merry way, you are not blameworthy. But it's not because you are exempt, nor is it because you failed to act. You act intentionally in stepping, and are responding to reasons (we may suppose), and the action is *yours*. But, you are not blameworthy because of your ignorance. Note that the typical way in which this works is by demonstrating that you did not display a malicious (or, say, negligent) quality of will.

The distinction between exemption and the two kinds of excuses ought to be fairly familiar from ordinary legal reasoning. I must be indicted before charges can be brought against me in a court (this analogous to finding that I am not exempt), and once I am there I can plead not guilty either on account of having not in fact done the thing in question (strong excuse), or on account of having done it in a non-culpable way (weak excuse).[15] Relevant to this, of course, is indeed the quality of my will. Whether I am guilty of malevolence, negligence, or excused altogether will depend on what I believed and what I desired at the time of my action.

We can now define two different kinds of responsibility. The first is:

Attributability: An action or state is attributable to an agent iff that agent is neither exempt from the sort of assessment appropriate for that action or state nor strongly excused from such assessment.

Attributability is a way of marking that at least two hurdles have been cleared: you are not exempt, and you are not strongly excused. If you step on my foot because of your blind spot, we can get at least this far. Blameworthiness goes further.

Blameworthiness: An agent is blameworthy for an action or state only if that state or action is attributable to her (she is not exempt from assessment and is not strongly excused) and is not weakly excused.

So, only if your action is attributable to you, and you have no excuses that justify the performance of it, is it appropriate for me to blame you. The kind of blameworthiness that I have in mind here is perhaps best thought of as a kind of liability. To be blameworthy in this sense is for a range of reactions of what we might call 'holding to account' to be appropriate.[16] These reactions include the Strawsonian reactive attitudes, such as resentment and the withholding of good will, but also things such as demands for compensation, material or otherwise. What unites

these reactions of holding to account is that they demand of the offending party a response, the appropriate response to which in turn is forgiveness. The simplest case of blaming someone in this sense is perhaps finding them blameworthy, demanding an apology and withholding good will until it is given. The appropriate response to a sincere apology is forgiveness and a repair of relations. The *compensatory* nature of the demands of accountability which are characteristic of this kind of blame thus distinguish it from punishment, which is retributive.[17]

My use of the term 'attributability' is closely related to that of David Shoemaker (2011) and Gary Watson (2004b). For Shoemaker and Watson, judgments of attributability are also grounds for aretaic, or characterological, assessments of agents. My use of the term is in accordance with their use in this respect. So, not only is attributability a logically necessary condition on blameworthiness, it is also in its own right typically grounds for a distinctive kind of assessment. When an agent 'owns' a state or action in the right way, it is expressive of her will in the sense that she thereby reveals to us something of her deep self: perhaps something about her desires and motivations; her perspective on life and on herself; or her characteristic patterns of thought, action, and evaluation.

This can be brought out in connection with the two different kinds of excuses. Since strong excuses work by undermining attributability itself, we should expect that when someone is strongly excused we find that there are no grounds for assessing him aretaically. And this is what we find. If you step on my foot because you were pushed, you do not thereby disclose yourself to me. On the other hand, if you are merely weakly excused you might not be an appropriate target for blame, but I may nevertheless learn that you are clumsy.[18] Sometimes, in addition to blocking the step from attributability to blameworthiness, a weak excuse will also provide grounds for a countervailing aretaic assessment. If I learn that you pushed me to save me from being hit by oncoming traffic, not only are you excused by demonstrating that the quality of your will was not malicious, you show yourself to be acting virtuously, in a way that merits praise. I will return to this function of weak excuses below.

I hope that it is reasonably clear how, according to my view of self-deception, self-deceivers are attributability-responsible for their self-deception, and that they are (typically at least) also blameworthy. The self-deceiver seems to violate an epistemic norm, and so we can begin anew an inquiry parallel to the questions asked when we inquired about whether you were blameworthy for stepping on my foot. Let us consider A. A has somehow come to the belief that his work is not flimsy. But it is manifest that this is not the case. He persists in his fantasy nevertheless. According to my view, this is because he omits to do what is necessary to bring his belief in line with the available evidence because he has a desire that his work not be flimsy. A is not exempt. (In general, it seems self-deceivers are not exempt; no creature without the capacities to be a candidate for moral assessment generally could be the subject of self-deception.) Is A strongly excused? Strong excuses work by showing that the action or state did not 'belong' to the agent in the right way, that it was not an expression of his will. Of course, it is possible for someone very much like A to act, and think, and speak like A and yet to be strongly excused. If A were being controlled remotely via a chip implanted in his brain perhaps he would be strongly excused. But as we are imagining him, A is engaged in a kind of fantasy which serves an important psychological function for him (though he is almost certainly unaware of it), and it reflects, on account of the motivation which my account attributes to him, his desire that things *be a certain way*, a way which they manifestly are not, and from which he has insulated himself. He thus is the owner of his self-deceptive omission(s), he does manifest his will in the process, and, importantly, he discloses himself and is an appropriate target for aretaic assessment.

Self-deceivers often elicit judgments of frustration, pity, and even contempt. These judgments first get a foothold at the level of attributability because they are appropriate responses to someone who has displayed the qualities of character that A has, viz., injudiciousness, vanity, and even cowardice. The only thing which remains to be determined is whether A might have a weak excuse that could insulate him from blame or potentially provide grounds for a countervailing aretaic assessment. But as far as we can tell—and as seems to be the case for self-deceivers quite generally—there is no excuse that A can appeal to. Weak excuses work by showing that the agent did not manifest a malicious or negligent quality of will, but A *does* manifest (at least) a negligent quality of will. Indeed, in self-deception we see the marriage of both epistemic and volitional defects combining to make for this negligence.[19] In willing something to be the case which is manifestly false, A both shows the epistemic vice of injudiciousness and is engaged in a flight from anxiety. This combination of epistemic and volitional failures strikes me as distinctive of motivated irrationality. Doing one's epistemic duty often requires a steadier will than the agent possesses, and this failure can manifest itself, on my view, as a motivated failure to seek, recognize, or appreciate evidence. Below I will discuss a case where a manifestation of epistemic vice seems to be excused, but that does not appear to be the case here. I now wish to turn to delusions.

Background: delusions

In this section I want to introduce delusions by way of a working definition, and by examples, many of which I will return to as we go along. By way of a definition, the DSM–V says (APA 2013):

> Delusions are fixed beliefs that are not amenable to change in light of conflicting evidence. Their content may include a variety of themes (e.g., persecutory, referential, somatic, religious, grandiose) […] Delusions are deemed bizarre if they are clearly implausible and not understandable to same-culture peers and do not derive from ordinary life experience […] The distinction between a delusion and a strongly held idea is sometimes difficult to make and depends in part on the degree of conviction with which the belief is held despite clear or reasonable contradictory evidence regarding its veracity.

Many parts of this definition are controversial, and it is substantially different from the DSM-IV version.[20] There is plenty to say about the definition and its relation to earlier ones but for now it suffices to note that the focus in the updated definition has shifted to what we might call the *epistemic* features of delusions. These features (fixity, degree of felt conviction, persistence in the face of clear contradictory evidence, etc.) are those that have most puzzled philosophers.

To get an idea for the variety of possible contents for delusions, here are some examples of (types of) delusions, individuated by their content:[21]

1. Delusions of persecution: Most common content for delusion (APA, 2000, p 299). The subject believes that his or her life is being interfered with from outside (almost but not always harmfully). Occurs in schizophrenia, affective psychosis, and in organic states.
2. Capgras delusion: Subject believes that a close friend or family member has been replaced by an impostor (Capgras and Rebould-Lachaux, 1923). I will return to Capgras in

below in connection with the 'two-factor' account (Davies, et al. 2001) of delusion formation and maintenance.

3. Anosognosia: The denial of illness. Often follows stroke or brain injury and involves denial of following disability, e.g., paralysis. Ramachandran's (Ramachandran, 1996) patient F.D. suffered a right hemisphere stroke causing left hemiplegia. But F.D. claimed she could walk and clap. Can also occur in schizophrenia, leading patients to refuse to take medication.
4. Reverse Othello delusion: Subject believes in the fidelity of his or her romantic partner in the face of strong evidence to the contrary. Peter Butler (2000) reports the case of B. X., who suffered a severe head injury in a high-speed car accident. Despite the absence of contact with his romantic partner, he subsequently 'developed an intense delusional belief that [she] remained sexually faithful and continued as his lover and life partner' (Butler, 2000, p 86). I will discuss B. X.'s case extensively below.

I should note just in passing that, despite the language in the DSM (and the language I have used here), it is a matter of some dispute amongst philosophers whether delusions should count as doxastic states. However, in what follows I will be assuming that delusions are best thought of as beliefs.[22]

Responsibility and delusion

Now that we have a working understanding of both self-deception and delusion on the table, and a sense of how self-deceivers are typically responsible for their self-deception on my view, our question becomes: are some delusional subjects self-deceived? Does self-deception play a role in forming and maintaining at least some delusional beliefs? Here I will argue that we should say 'yes'. This may seem surprising not least of all because self-deception typically concerns matters which are much more 'garden variety' than the bizarre contents of delusional belief, however, not all delusions have such bizarre content, as we shall see. And even where the content is bizarre, there is room for motivation to be playing a role that might imply self-deception is at work.

If my account of self-deception is correct, it seems to provide relatively straightforward criteria for assessing whether self-deception is implicated in delusional belief. We must only ask whether it is true that the agent has failed to confront, for motivationally biased reasons, manifestly available evidence that would overturn her belief. What remains, however, are two tasks which are not so straightforward: first, we must try to determine whether any actual delusions satisfy those criteria, and further, we must determine what kind of responsibility, if any, that would ground. Let us first address head-on the question of whether any delusions can be thought to fit my model of self-deception. The most plausible candidate for such a case is the Reverse Othello delusion.[23]

Reverse Othello delusion. Recall Peter Butler's patient from before, B.X.. B.X. suffered severe head injuries in a car accident. As a result of the crash he was left quadriplegic and unable to speak without the use of an electronic communicator. According to Butler, in the initial stages of his illness he expressed both insight and 'intense emotional response to a massive disability and a fracturing of his interpersonal relationships' (2000, p 87). However, in the year following his injury, B.X. gradually developed the delusional belief that he was still in a successful romantic relationship with his former partner (who left him following his injuries) and even claimed that they had recently married, occasionally claiming that he needed to leave treatment to return home to his wife.

B.X.'s appreciation of his injuries is important. He is trying to come to terms with the significance of an irreversible life-changing calamity, and seems to be doing it head-on. But there is a limit to how much such change he can accept at once without falling apart. Butler characterizes B.X.'s delusion as protecting him from falling into severe depression, or as we might say, existential collapse. For him, the ability *to go on* is contingent on his believing that his former partner remains faithful to him. To lose her, on top of all of that has already happened, would be, in some sense approaching the literal, unbearable.

In this context it is also important to note that B.X. eventually manages to recover from his delusion. Even when the delusion was at its most elaborated B.X. did not experience any other psychotic symptoms. The delusional belief seemed to dawn on him somewhat gradually, and eventually reached its most elaborated form in the idea that he and his former partner had been recently married. But the delusion also gradually receded, and he came to accept that she had no intention of returning to him. It is as though the delusion held at bay the need to face something that B.X. was not capable of accepting, until such time as he was more fit to do so.

Together these two things suggest that B.X. is reasons-responsive generally. His initial sensitivity and insight into his condition are not things that he could have displayed if he had crossed that strange boundary that leads outside the space of reasons altogether. And the fact that he was able to recover more or less on his own suggests that his capacity to be sensitive to epistemic reasons remained intact—for what else other than that very capacity could he have used to get himself out?—even if it suffered partial muting and redirection.

On Butler's way of thinking of things—to which I am obviously very sympathetic—B.X.'s belief served a protective or defensive function. How did it serve that function? It is plausible that B.X. needed (in some appropriate sense) to believe something which would forge a strong sense of coherence and connection with his pre-injury self. As Butler suggests, the primary challenge for B.X. during his recovery was coming to terms with how dramatically and irreversibly changed his life had become. Believing that his partner was there, that a dear corner of his otherwise unrecognizably marred life remained as before, could plausibly offer him something to hang on to, some piece of his past life to use as a flotation device while he tries to get himself to shore.

Now, according to my view of self-deception, it seems that B.X. counts as self-deceived. The belief that his partner had not left him made its appearance sometime after the period of insight that Butler describes. This suggests that B.X. had the belief that his partner had indeed left him at some point prior to the onset of the delusion. Now let us suppose that as the significance of how his life has been transformed dawns on him bit-by-bit B.X. develops a need to believe that his partner had not left him. In a number of ways such a belief is a good candidate for a life-preserver-belief because it concerns a matter which is indeed of great personal significance for him but, compared to the other things of great personal significance to him which are manifestly in shambles it less often and less flagrantly bumps up against evidence to its contrary which would need to be ignored in order for the belief to persist.

It seems that if B.X. is to persist in his life-preserver belief he will need to avoid confronting evidence which points to its falsity. His case is an interesting one because presumably the evidence which is available concerning the falsity of that belief is given by his memory and the memories of his caretakers. Compared to the body of evidence that would have to be ignored if he were to, say, try to deny his injuries (as some delusional patients do), this body of evidence is quite sparse. A little bit of motivated failure to consult that part of one's memory might be all that would be

required. If this is what happened, then B.X. counts as self-deceived according to my view. His belief, however it was formed precisely, is false, and manifestly so. But he manages to persist in believing for a time (as long as he needed to, it seems) and this seems to require making himself somehow impervious to the evidence which he had previously appreciated.

The possible complicity of his caretakers in facilitating his failure to confront or appreciate evidence against his delusional belief is another interesting feature of B.X.'s case. It is also quite readily understandable. Clinicians often have to face the difficult question of whether it is appropriate to confront a subject about their delusional belief and many factors might go into determining the appropriate course of action. Plausibly, it would have seemed to many clinicians that the right course of action in B.X.'s case would be to allow him, as he seemed himself to want to do, to take things one at a time, so to speak. If the self-deceiver is encouraged or facilitated in their motivated omission, I am inclined to think that it may partially mitigate his degree of blameworthiness.

Although I do think B.X. is self-deceived on my view, I do not think he is blameworthy. A very small part of the reason for this might be the facilitation of his caretakers. But far more important, it seems to me, was the function that the self-deceptive/delusional belief was playing for B.X. at the time. If it really was (perhaps the only thing) keeping him together, then I think we are right to see it as an excusable trade-off between negative epistemic value and significant improvement to overall well-being. This does not change the facts concerning *whether* B.X. in fact deceived himself, but it is certainly relevant for determining what the appropriate attitude is to take towards him in the light of his self-deception. I chose to express this by saying that the self-deceptive omission, and the resulting delusional belief are attributable to B.X., but that he is not blameworthy for the subsequently persistent belief because he has a weak excuse.

What of the aretaic assessment that is typically grounded by attributability-responsibility? Is B.X. injudicious in the same way that A is? Does he display a negligent quality of will? He may. The moral hazards of self-deception—risk of harm to self or others, for example—are there just as much in his case as in others. But the weak excuse that is available in B.X.'s does more than just block his blameworthiness. It also provides ground for counter-balancing, or perhaps undermining, the aretaic assessment that would normally apply.[24] Excuses of this kind work as follows. Suppose I am tasked with delivering some valuable cargo. If, on the way down the only available path, I encounter a hairy spider and decide to turn back, risking the cargo in the process, I am pretty clearly guilty of cowardice. The action of fleeing is attributable to me (and is the grounds for finding me cowardly), and so too am I blameworthy (liable) if the cargo is lost. On the other hand, if I turn back risking the cargo because there is a grizzly bear on the path, I do not display cowardice, but perhaps prudence. (Or, if one prefers, I display cowardice tempered with prudence.) For the same reason that the negative aretaic assessment of me would seem inappropriate, I submit that I am not blameworthy should the cargo end up lost; the very thing that excuses me from blameworthiness also undermines or counter-balances the judgment of cowardice. This seems to be what is happening in B. X.'s case. His flight from the truth is analogous to my flight from the bear: it comes with risks that we all recognize, but it is not undertaken lightly or negligently. The quality of will that he displays, against the situation in which he finds himself at the same time makes blaming him, and finding aretaic fault, inappropriate.

It is worth mentioning that I want to resist saying that B.X. has a strong excuse for his self-deception. To say this would be to deny B.X. the appropriate ownership over the strategy that he deployed for getting through. I have spoken of his psychological *need* to believe as he did, but I did not mean to suggest that his deceiving himself was something he was literally compelled to do. And more importantly, merely being compelled to do something in this sense might not be enough to constitute a strong excuse. I said that if you only stepped on my foot as a result of being pushed, you would have a strong excuse because the action would not be yours. What if you also, simultaneously *wanted* to step on my foot? Then your action would have a cause outside of you, but would also be an expression of your will. If we were in this situation, we would have to do hard work to figure out which thing should properly be considered the reason for your bodily movement. I do not have a general procedure for coming to answer questions of this kind, but I think it is safe to say here that B.X. is not overdetermined in this way. It is clear that what he does is a manifestation of his will, even if there is a sense in which he must do it, because he wills to do as he must. This would not be true of you if you were both shoved and malevolent; your will was not to be shoved, even though it may have been to step on my foot. Being compelled may not be enough to constitute a strong excuse if one also wills the means.

Capgras; two-factor theory. While I do think that B.X counts as self-deceived on my view, it is unclear how many other cases of delusional belief will satisfy my account of self-deception. My aim has certainly not been to argue that all delusional subjects are self-deceived, or even that it is the norm. However, I think it is worth pointing out that my approach here dovetails quite nicely with a prominent approach in cognitive neuropsychiatry to the formation and maintenance of delusion which is called the 'two-factor theory', and which I mentioned earlier in connection with Capgras. This raises the possibility that motivational factors akin to those that I think are at work in self-deception may be at work in more cases than is widely recognized. Let me elaborate.

Recall the Capgras delusion. Someone with this delusion believes that a friend or family member has been replaced by an impostor. Understanding of how this delusion is formed was greatly enhanced by the discovery that the human facial recognition system has at least two neurologically independent subparts. The first, which is responsible for 'overt' facial recognition is in the temporal lobe and underlies the ability to explicitly recognize the faces of those one is familiar with. The second, affective, system, which appears to involve the amygdala, produces Skin Conductance Responses (SCRs)—*covert* recognition—when subjects are exposed to faces they are familiar with, even if they fail to recognize the face overtly. This is what is thought to be at work in people who have prosopagnosia.

This insight led cognitive neuropsychiatrists studying Capgras to wonder whether the two facial recognition systems were doubly dissociated—that is, whether each was independent of the other and whether there could be people who had the 'opposite' of prosopagnosia. Such people would overtly recognize familiar faces, but would be left without the typical accompanying affective response. Could this be what was causing Capgras? The patient would see his wife, and would accept that the person before him bore an exact physical resemblance to her, but the experience would be entirely without the ordinary feeling of familiarity. It is, perhaps, only a small leap from there to the idea that this person before me, while she looks exactly like my wife, must be someone else.

The two components of the facial recognition system are now largely thought to be doubly dissociated and the abnormal experience of seeing someone who looks exactly like a loved one, but who feels somehow alien, is thought to be involved in Capgras (Ellis and Young, 1990; Ellis et al. 1997). There is, however, a

problem. Not everyone with damage to the covert facial recognition system develops the delusion. Even though these patients are having the same unusual experience as the Capgras patients, they do not form the delusion. So, something else must be required to fill in the gap between the unusual experience and the subject eventually forming or endorsing the belief. This leaves somewhat unsettled what the second factor must be (and there is no consensus) but the idea that some kind of two-factor theory is correct, at least for some delusions like Capgras, seems difficult to deny given the evidence. There must be some role for the abnormal experience to be playing, but if that does not take us all of the way there, there simply must be something else at work.

Many of those who pioneered the two-factor theory were responding to an idea, tracing back to the work of Brendan Maher beginning in the 70s,[25] that delusions were largely rational responses to highly unusual experiences (Davies, et al. 2001; Davies, et al. 2010). Maher himself thought of his work as a direct challenge to Jaspers' claim that delusions were un-understandable. If some delusions could be understood as rational responses to a certain special kind of experience, experience with a certain kind of force and character, then the content of the beliefs that those experiences gave rise to could be readily understood.

Whether this rational connection can be maintained, and what it means for self-deception depends on how we think of the relation between the first and the second factor. One way of getting at this is to ask how specific the representational content of the abnormal experience is. For example, according to Coltheart (2005), the Capgras patient does not *experience* that his wife is an impostor; rather, an unconscious system predicts that seeing his wife should be accompanied by a certain autonomic response which fails to occur. He thus forms the Capgras hypothesis as an attempt to *explain* the abnormal experience. According to this 'explanationist' account, the representational content of the experience which prompts the delusion is less rich than the content of the delusion itself. According to the competing 'endorsement' account (Bayne and Pacherie, 2004), the representational content of the experience prompting the delusion is as rich as the content of the delusional state itself. On this kind of account, the subject does not reach for the Capgras hypothesis as an explanation of his experience but merely takes what is already presented in experience to be veridical.

Obviously, whether there is room for appeal to motivational factors (and whether such an appeal would make a given case count as self-deception on my view), depends on which of these competing accounts is true. On either account, motivational factors (possibly jointly with neuropsychological factors) could be playing a role in generating the anomalous experience. Since the experience is much thinner on the explanationist model, it might be thought that appeal to motivation would be otiose; still, there could be a role for it to play. If, as some philosophers—and increasingly many psychologists—think, it is possible for a subject's propositional attitudes to *cognitively penetrate* (Pylyshyn, 1999) her experience, then two subjects may have different experiences even if we hold fixed what is perceived, the perceiving conditions, and the state of the relevant sensory organ. If this is right, then the mere fact that one subject desires that *p* be the case while the other fails to desire it or desires that not-*p* be the case, might just be the difference-maker when it comes to answering the question, 'Why did the subject have the experience that he had?'—even on the explanationist model.[26]

If motivation is playing a role in the first factor, that will not be enough for the sufficient condition identified by Self-deception as Omission to be satisfied. When we learn that the subject had some distinctive kind of experience, we just have not learned one way or another whether there has been motivated mismanagement of evidence which sustains an externally defeated belief over time. However, we may nevertheless be able to learn something about the subject that undergoes such an experience which is akin to what we can learn about the self-deceiver. If we are interpreters of someone who has undergone an experience of this kind, and we learn that this is how it has happened, we come to learn something about the kind of cognitive agent that the subject is.

There is also room for motivation to be playing a role in the second factor,[27] and if it is present, it may go some of the way to restoring the kind of understandability that Maher was aiming for. On either the endorsement or the explanationist account of things, if we have gotten this far, the Capgras belief is already in place, either as an explanation for a bizarre experience or as one given rise to by a bizarre experience directly. Once the belief is in place, there is room for Self-deception as Omission to be satisfied. All that would need to be the case would be for there to be a failure of epistemic agency which is partially motivated by a desire for the world to be as the subject already believes it to be.[28] And as strange as it may sound, the operation of the second factor seems more readily understandable when it is cashed out in motivational terms, or indeed in terms of the kind of mental agency that I think is at work in self-deception. The varieties of human motivation are nearly limitless, and I do not know of any clinical examples that bear this out, but it is not difficult to imagine someone facilitating the maintenance of the Capgras belief for motivationally biased reasons.[29] Perhaps the couple has recently had a particularly acrimonious quarrel and it would be somehow easier to not face the genuine article just yet; perhaps he has been secretly yearning for a divorce and this would save him the trouble; perhaps he has a motivation which only years of deep analysis would uncover. Any such motivation, if it were to underlie and facilitate the acceptance of the Capgras hypothesis, would be grounds for thinking that we had a potential case of self-deception here. The availability of an explanation of this kind greatly reduces the sense that the delusion is un-understandable by bringing the psychological dynamics of the subject into the focus of ordinary intentional explanation.

Conclusion: innocence

I have tried to bring a number of distinctions between types of responsibility to bear on the question with which we began. I have also put the notion of self-deception to work in a way that I hope has been doubly illuminating: since we have defeasible but determinate antecedent judgments about the responsibility-status of self-deceivers, asking whether someone's conduct can be assimilated to a self-deceptive paradigm can help us think about the ways in which they may or may not be responsible. Delusions can also help us understand the ways in which our ordinary notion of self-deception can be extended to include, e.g., cases where the self-deception is attributable but not blameworthy for very good reason. Using self-deception as a tool for thinking about some delusions also forces on us the question of what a subject's *motivations* are and this question can only be answered by (suitably supplemented) ordinary interpersonal interpretation. Motivations can partially constitute nodes of intentional agency and reminding ourselves about the motivations of subjects with delusions and the role that such motivations may play in our assessment of them can serve as a general bulwark against slipping too easily into thinking of them as outside of the scope of ordinary assessment and understanding altogether.

It is telling that when we bring to bear the tools that I have recommended for thinking about responsibility for delusions we find that there is a good case to be made that the subject is excused. I take this to be in keeping with something Lisa Bortolotti has

recently argued for (2015, 2016), viz., what she calls the 'epistemic innocence' of some delusions. She says that a delusion is epistemically innocent if it confers significant epistemic benefits which could not be achieved otherwise. Bortolotti is focused on cases where some negative epistemic consequences are embraced for the sake of otherwise unattainable *epistemic* benefits. I agree with Bortolotti that the notion of epistemic innocence is of clinical and conceptual value. What I hope to have done here is to have introduced what might be thought of as an expansion of that notion of innocence to cases where the negative epistemic consequences are traded off against *non*-epistemic gains. In order to address such cases we need conceptual tools developed from the more general standpoint of moral theory. Taking up this standpoint—taking seriously the possibility that assessment might here really be appropriate—has, I hope, revealed a more comprehensive and detailed picture of what that innocence consists in. There is a kind of innocence which may only be possible against a backdrop of possible guilt.

Received: 14 October 2016 Accepted: 2 October 2017
Published online: 31 October 2017

Notes

1 But only roughly. I explain how my use of this term—and the distinction I use it to mark—differs from Shoemaker's below.

2 Foucault located such dynamics within what he called 'discourses' which are 'ways of constituting knowledge, together with the social practices, forms of subjectivity and power relations which inhere in such knowledges and relations between them' (Weedon, 1987, p 108). I do not wish to problematize the knowledge that discourses enable (as Foucault did) but merely to draw attention to the tacitly normative aspects of certain practices of categorization.

3 This example is adapted from Doggett (2012).

4 Note that the view is stated as a sufficient condition. Because the satisfaction of this condition can be sufficient for something to count as self-deception, the constitutive connection thesis (see below) must be false. But that is not to say that there *could not* in principle be cases of self-deception which do not satisfy the condition given here. (It would be a separate question, however, whether such cases and the view meant to capture them would run afoul of the dynamical problem.)

5 I use the terms 'satisfaction' and 'dissatisfaction' deliberately to avoid commitment to the idea that there must be positively or negatively valenced *experiences* accompanying the subject's belief. However, there no doubt will be cases where the subject will *experience* satisfaction or believing p or will experience something like *distress*.

6 Theorists of self-deception disagree about what form the subject's emotional interest must take. For Mele (2006, 1997), for example, what it is to have an emotional interest in p's being true is to take the error of mistakenly believing not-p when p is in fact the case to be more costly than the error of taking p to be the case when in fact not-p is the case, where that preference is itself to be understood in terms of a motivational bias. So, on this view, I might prefer to believe that I am not going bald to believing that I am going bald because believing that I am going bald if I am not would cause me great distress, much more than the distress that I would feel if I mistakenly believed that I was not going bald when I was. Barnes is more explicit about self-deception's anxiety-reduction function. She says (1997, p 39): 'When a person is anxious that not-q, the person (1) is uncertain whether q or not-q and (2) desires that q'. Self-deception reduces the person's anxiety by resolving the question of whether q in the appropriate direction. Although my formulation is much closer to Barnes', as far as I can tell my use of the idea of 'desiring that p be true' is consistent with both of these ways of thinking about self-deceptive motivation. That is, having the motivated error-preferences that Mele is pointing to is as much a matter of being emotionally entangled (in my sense) as anxiously desiring that p be true.

7 It is also worth noting that both Barnes and Mele are keen to be able to handle cases of 'twisted' self-deception, i.e., cases where the subject self-deceptively believes something he wants *not* to be true. My way of putting things can also handle these cases once we distinguish between wanting something to be true in the ordinary sense and desiring it to be true in the sense that I mean it here. A subject may desire (in my sense) for something to be true (we may say) *masochistically*. That is, he may be liable to satisfaction at believing that p where p is something that is bad for him, something he, in the ordinary sense, does not want (or wants the opposite of). Believing that p closes the gap between the way he takes the world to be and the way he thinks the world *ought to be*, but the way that the world ought to be from his point of view is bad (say, hedonically bad) for him.

8 I add this qualification to avoid having to individuate processes, and will omit it from here on.

9 Modulo, should there be such a thing, natural credence extinction.

10 I thank an anonymous referee for formulating this worry in this way to me.

11 If Mele would be willing to understand his view as including cases of culpable belief maintenance, our views would be very closely related.

12 I am not sure that Mele succeeds in avoiding the dynamical problem, which is part of the reason why I think we need to go further and deny the constitutive connection thesis, that is, move to talking not about belief formation but about the dynamics of belief maintenance. For critical discussion of Mele and the dynamical problem see Lockie (2003).

13 There is thus an affinity between my view and Fingarette's (1969).

14 Using the term 'excuse' to refer to a specific variety of what Strawson (1962, p 5) called 'pleas' or 'special considerations' is due to Watson (2004b, p 224). I take this way of distinguishing between two types of excuses to be intuitive, however it has not, to my knowledge, been drawn in this way and in connection with the two types of responsibility I wish to distinguish.

15 The analogy is limited in the following way: There is difference between what is standardly called 'excuse' and justification. If an agent can produce either, she can be shown to have avoided culpable wrongdoing. An action is justified if, all things considered, it was not wrong. (self-defense); an agent is excused if something undermines her responsibility (acting under hypnosis or duress). Either could be presented as a defense against criminal charges, but weak excuses, as I intend them to be, only comprise the latter.

16 It is worth noting, however, that whereas Shoemaker contrasts attributability with what he calls 'accountability' (as well as a third notion, 'answerability'), my understanding of blameworthiness should not be identified with Shoemakerian accountability. For Shoemaker, accountability has specifically do with violating relationship-defining demands (which play no role in my discussion).

17 So there ought still to be a considerable gap between someone's being blameworthy, and the truth of the claim that we should actually *punish* them. Even if someone is found fully blameworthy for an action attributable to them it may still be a wide open question what the right kind of response to their wrongdoing is. Indeed, it is compatible with my way of thinking about moral responsibility that punishment is seldom, if ever, justified or appropriate. We can see this if we imagine adopting a flat-footed consequentialist justification for punishment. What if it turned out that punishment was an ineffective deterrent and a poor means of personal rehabilitation? It would not follow from this that no one's actions were ever properly attributable to them, nor that they were not blameworthy for those which were bad. It would simply mean that punishment would not be the response justified by those facts. And it seems immensely plausible to me that if anything would make punishment inappropriate, it would be the fact that the wrongdoer already suffers great enough misfortune that further punishment would take on a perverse character. This is just to say that the question guiding this paper is certainly not the question of whether some delusional subjects should be *punished* for their delusions.

18 Whether you are excused for being *that way* (or whether it is countervailed by another aretaic assessment) is yet another question which will arise again in connection with delusional subjects below.

19 There are some habits of mind which are epistemic vices only because they are accompanied by indolence. For example, perhaps all of us are subject to the availability heuristic, or liable to commit the base rate fallacy, or to manifest various other System 1 cold biases. What separates those of us who allow the errors characteristic of those biases to persistently take hold and those who do not is some degree of epistemic vigilance, which is effortful. For example, consider the following example of Kahneman's to illustrate the System 1 at work. He says: 'Do not try to solve it, but listen to your intuition' (Kahneman, 2011, p 44):

> A bat and a ball cost $1.10.
> The bat costs one dollar more than the ball.
> How much does the ball cost?

The intuitive—incorrect —answer that System 1 offers up is $0.10. And it seems to do so more or less unbidden, once the specification of task has been grasped. Whether one chooses to go on and perform the calculation and ultimately arrive at the correct answer seems to be an independent matter. Without the exercise of vigilance, one is saddled with a false belief. This is not an example of self-deception, but it is a nice illustration of how some epistemic vices are enabled by unwillingness. In cases where this is true, there is a foothold for various forms of assessment, including aretaic assessment. (I leave open the question of whether there are cases of 'pure' cognitive bias and what kind of assessment, if any, would be appropriate there.)

20 In particular, the requirement that the belief be false, that it be based on 'incorrect inference', and that it be bizarre, have all been weakened or dropped. These ought to strike us as welcome changes.

21 Delusions are typically differentiated by their content, but it is also widely acknowledged that delusions are partially dependent on their cultural milieu for the particular contents that they have (Ahmed, 1978).

22 One reason is simple: I am very sympathetic to the idea that belief is at bottom a concept we use to understand other agents, to explain and predict their behavior in terms that are readily understandable to us, and generally to calibrate them in the space of reasons. To say that someone believes *p* might mean things as various as (i) they are inclined to act on *p*; (ii) they are inclined to report *p* in speech; (iii) they are inclined to use *p* as a fixed point in practical or theoretical reasoning; (iv) they have a certain felt conviction in the truth of *p*; (v) they treat the question of whether *p* to be largely settled etc. (Scanlon, 1998). So, on this way of thinking about it, belief is a kind of syndrome with no essential features, and someone can be thought to count as believing that *p* by exhibiting some number of the marks of beliefs. My sympathy with the idea that delusions ought to count as beliefs stems largely from the incontrovertible way in which delusional subjects satisfy, albeit in shifting and sometimes patchy ways, these criteria. In particular, it is very difficult to deny that patients with delusions *take themselves* to believe the things in question. They are subjectively experienced as ordinary beliefs and indeed, one's degree of felt conviction in a delusion can often greatly exceed the conviction one might experience in ordinary belief. As Sims (2003, pp 141–142) puts it:

'It cannot be stressed too often that patients believe their delusions literally: subjectively, delusions are completely different from fantasy. Patients do not describe them 'as if' they existed. The reality is 'known' with the unconcerned certainty that the undeluded person assumes for the concrete events and ideas of his own life, such as the floor being solid…[A] man who believed that American battleships were sailing down the main street of Birmingham UK (100 miles from the sea), had the refined social conscience to report this to the police!'

This example is a nice illustration of how delusions may exhibit some of the marks of belief with clarity and sharpness, even while exhibiting many of the negative epistemic features which are characteristic of them. This subject is using his belief that there are battleships sailing down the main street of Birmingham as the basis for speech, inference, and indeed concern (i, ii, iii) *because* of his degree of felt conviction (iv) in it. Moreover, it seems that the very fact that he takes such a thing to be a reason for concern shows that his belief exhibits a degree of coherence with his other beliefs, beliefs, e.g., about geopolitics and nationhood (not to mention a whole lot of beliefs about military hardware, the nature of peacetime etc.) which together suggest that what is happening is cause for some alarm. The belief is no doubt implausible, and we can imagine that it exhibits a high degree of fixity and resistance to counterevidence, but that should not disqualify it from counting as belief.

23 The Reverse Othello delusion is noteworthy among delusions for not having the same kind of bizarre content that most delusions have. It may in this respect seem tailor-made for someone who wants to defend the claim that there is overlap between self-deception and delusion. A critic might say: 'Most delusions involve believing a highly bizarre content, and it is plausible that (part of) what is distinctive about being in that state is *how* the subject comes to have that attitude towards that content (perhaps, e.g., it is caused by unusual perceptual experience.) But then there will be a gap between self-deception and delusion, one that is missed if we focus only of belief maintenance.' I thank an anonymous reviewer for raising this issue to me. My response is twofold: First, trying to figure out how delusions are formed is surely to be counted as one of the chief aims of the neuropsychology of delusions. And it is very plausible that for a great many of them there will be abnormal mechanisms at work that partially explain (among other things, perhaps) the bizarreness of the delusional content. And this may mean that there is, in some respect, a gap between (some) delusions and (some) self-deception. But the size of this gap, and its significance, will depend, in part, on what the neuropsychological abnormalities in question turn out to be like. We can see this by considering how, when we think of the neuropsychological abnormalities that we *do* know about already, it is still just a good question the extent to which this shows there to be 'gap' of relevance to thinks like, say, our responsibility judgments. When this-or-that neuropsychological abnormality is discovered, it will still be a good question the extent to which that abnormality presents or underlies a philosophically interesting discontinuity with ordinary cognition, agency, autonomy, or whatever. Second, even if it turns out that the relevant abnormalities do underlie significant discontinuities between the delusional and the non-delusional with respect to belief formation, it may remain true that there is an interesting overlap between self-deception and delusion precisely because the mechanisms of belief formation are not the only ones we must look to if we want a comprehensive picture of the ways delusional subjects are and are not like 'ordinary' subjects.

24 I make this qualification because I want to remain neutral with respect to whether the virtues are necessarily unified.

25 Such as, e.g., Maher (1974).

26 Of course, this could also be the case on the endorsement model. Indeed, I am assuming that the connection between what the subject desires and the content of the experience would be easier to see on this model since the content of the experience is identical to the content of the delusional belief. So, whenever it is plausible that the subject could desire that the delusional belief be true, it will be plausible that the subject desire that the content which shows up in his experience be true. (Of course, whether cognitive penetration works this way—whether desiring that *p*, say, probabilifies an experience with the content that *p*—is an empirical matter.)

27 Davies (2010) also discusses cases where motivational factors could be playing a role between the first factor and the second.

28 Mele denies that self-deception obtains when the subject also suffers from a cognitive impairment, on the grounds that the 'causal contribution [of motivation] may be so small' (2006, p 123) that it should not count. I do not see why we should say that it does not count rather than say that the contribution that it makes is gradable. And even in cases where (as Mele has in mind) the reflection on the available evidence that is prevented by motivational factors would not have caused the subject to revise his beliefs, it seems to me that there is a characterologically relevant difference between the agent whose reflection is prevented by motivational factors and the agent whose reflection is not, one that we may well register by calling the former self-deceived and not the latter. When we ask whether someone is self-deceived, we are not *just* asking which factors are *causally* responsible for sustaining his beliefs. We are asking whether he manifests a certain epistemic vice.

29 I was told anecdotally of a case where a patient had stopped her medication in an attempt to manage her symptoms without the distressing side-effects that the medication caused. After it became clear that her symptoms were unmanageable without assistance her psychiatrist recommended, to her great dismay, that she restart the medication, to which she responded 'You're not Dr. X! He would never treat me this way!' This suggests that there are cases of patients forming the Capgras delusion *without* any underlying bizarre perceptual experience and where motivational factors seem to be doing most of the work. Conversely, there have been cases reported of people experiencing the Capgras delusion with their pets, or with inanimate objects (Islam, et al 2015.), where motivational explanations seem far less plausible. The size of the gap that needs to be closed by the second factor is evidently highly variable, as is the force of the motivational component. But this does not show that motivational factors are not playing a role in some cases of Capgras, it just shows that the role played may be greater or lesser (or perhaps nil) depending on the case.

References

Ahmed SH (1978) Cultural influences on delusion. Psychiatr Clin 11(1):1–9

American Psychiatric Association (2000) Diagnostic and statistical manual of mental disorders, 4th edn. American Psychiatric Association, Washington DC, Text Revision

American Psychiatric Association (2013) Diagnostic and statistical manual of mental disorders, 5th edn. American Psychiatric Association, Washington DC

Barnes A (1997) Seeing through self-deception. Cambridge University Press, Cambridge, UK

Bayne T, Pacherie E (2004) Bottom-up or top-down? Campbell's rationalist account of monothematic delusions. Philos Psychiatr Psychol 11:1–11

Bortolotti L (2015) The epistemic innocence of motivated delusions. Conscious Cogn 33:490–499

Bortolotti L (2016) Epistemic benefits of elaborated and systematized delusions in schizophrenia. Br J Philos Sci 67:879–900

Butler P (2000) Reverse Othello syndrome subsequent to traumatic brain injury. Psychiatry 63:85–92

Capgras J, Reboul-Lachaux J (1923) "Illusion des sosies" dans un délire systématizé chronique. Bulletin de la Société Clinique de Médecine Mentale 2:6–16

Coltheart M (2005) Conscious experience and delusional belief. Philos Psychiatr Psychol 12:153–157

D'Cruz, J (In prep) Unwitting pretense and the self-deceptive mind

Darwall S (1988) Self-deception, autonomy, and moral constitution. In: Mclaughlin B and Rorty AO (eds) Perspectives on self-deception. University of California Press, Berkeley, CA, p 407–430

Davidson D (ed) (2004) Deception and division. In: Problems of rationality. Oxford University Press, Oxford, UK, p 200–210

Davies M (2010) Delusion and motivationally biased belief: self-deception in the two-factor framework. In: Bayne T and Fernández J (eds) Delusion and Self-deception: affective and motivational influences on belief formation. Psychology Press, Hove, UK, p 71–86

Davies AMA, Davies M, Ogden JA, Smithson M, White RC (2010) Cognitive and motivational factors in anosognosia. In: Bayne T and Fernández J (eds) Delusion and self-deception: affective and motivational Influences on belief formation. Psychology Press, p 187–225

Davies M, Coltheart M, Langdon R, Breen N (2001) Monothematic delusions: Toward a two-factor account. Philos Psychiatr Psychol 8(1/2):133–158

Doggett T (2012) Some questions for Tamar Gendler. Analysis 72:764–774

Ellis HD, Young AW (1990) Accounting for delusional misidentifications. Br J Psychiatr 157:239–248

Ellis et al. (1997) Reduced autonomic responses to faces in Capgras delusion. Proc R Soci Biol Sci 264(1384):1085–1092

Fingarette H (1969) Self-deception. Humanities Press, New York, NY

Foucault M (1969) L'archéologie du savoir. (The Archaeology of Knowledge, Allan Sheridan, Trans.) Harper and Row, New York, NY

Gendler T (2007) Self-deception as pretense. Philosophical Perspectives 21 (1):231–258

Islam L et al. (2015) Cagras delusion for animals and inanimate objects in Parkinson's Disease: a case report. BMC Psychiatry 15:73
Jaspers K (2007) Causal and 'understandable': relationships between events and psychosis in dementia praecox (schizophrenia). In: Sass H (ed.) Anthology of German Psychiatric Texts, Blackwell Publishing, Oxford, pp 174–279
Kahneman D (2011) Thinking fast and slow. Farrar, Straus, and Giroux, New York, NY
Kendler KS, Campbell J (2014) Expanding the domain of the understandable in psychiatric illness: An updating of the Jasperian framework for explanation and understanding. Psychol Med 44:1–7
Lockie R (2003) Depth psychology and self-deception. Philos Psychol 16 (1):127–148
Mackay R, Kinsbourne M (2009) Confabulation, delusion, and anosognosia: Motivational factors and false claims. Cognit Neuropsychiatr 15(1/2/3): 288–318
Maher B (1974) Delusional thinking and perceptual disorder. J Individ Psychol 30:98–113
Mele A (1997) Real self-deception. Behav Brain Sci 20(1):91–102
Mele A (2006) Self-deception and delusions. Eur J Analytic Philos 2(1):109–124
Pears D (1984) Motivated irrationality. Oxford University Press, Oxford, UK
Pylyshyn Z (1999) Is vision continuous with cognition? Behav Brain Sci 22:341–365
Ramachandran VS (1996) The Evolutionary biology of self-deception, laughing, dreaming, and depression: Some clues from anosognosia. Med Hypotheses 47 (5):602–632
Scanlon T (1998) What we owe to each other. Belknap Press of Harvard University Press, Cambridge, MA
Shoemaker D (2011) Attributability, answerability, accountability. Ethics 121 (3):602–632
Sims A (2003) Symptoms in the mind: An introduction to descriptive psychopathology, 3rd edn. Elsevier Science, Saunders
Strawson P (1962) Freedom and resentment. Proc Br Acad 48:1–25
Watson G (ed) (2004a) Responsibility and the limits of evil: Variations on a Strawsonian theme. In: Agency and answerability, Clarendon, Oxford, p 219–259
Watson G (2004b) Two faces of responsibility. In: Watson G (ed) Agency and answerability, Clarendon, Oxford, pp 260–288
Weedon C (1987) Feminist practice and poststructuralist theory. Blackwell, Cambridge, 1987

Data availability

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

Additional information

Competing interests: The author declares no competing financial interests.

Reprints and permission information is available online at http://www.nature.com/reprints

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.