AUTONOMY AND FALSE BELIEFS

For those who believe in the possibility of autonomy, it is generally accepted that it is better for an agent, ceteris paribus, if any given action is performed autonomously than if it is performed non-autonomously.1 Such an intuition, I take it, lies behind the continuing interest in establishing the conditions for autonomous action. Since Harry Frankfurt (1971) first proposed his second-order desire model, according to which the desire that moves us to action must be validated by a higher order desire, moral philosophers have sought to explain what differentiates autonomous from non-autonomous action. The majority of this attention has focused on what I will be broadly classifying as reflection. The quality of an agent’s reflection on her potential action (or motivating desire or value) is taken to determine whether or not that action is autonomous. So it has been suggested that an action is autonomous if the agent would not feel alienated from the desire that motivates it, were she to reflect on the historical origins of that desire (Christman 1991, 2009). Alternatively, it has been suggested that an action is autonomous if it is performed on the basis of a preference that has been vetted by an agent’s conception of the good (Ekstrom 1993). Raising the stakes somewhat, it has also been suggested that an action is autonomous only if it is motivated by norms that are not false and oppressive (Stoljar 2000). Though very much at odds with one another, what all of these accounts of autonomy share is that they situate the conditions for autonomy in the quality of reflection (hypothetical or otherwise) that the agent undertakes with respect to her potential action.

In this paper, I argue that there is something missing in all accounts of this nature. By focusing overwhelmingly on the way in which the agent reflects, such accounts overlook the importance of what the agent is reflecting upon. Whichever of the current formulations of autonomy we accept (and for the purposes of this paper I remain for the most part agnostic between them), reflection could be undertaken in full accordance with the conditions set, and yet the action fail to be autonomous. This will occur, I argue, if the agent is mistaken about the object of her reflection. More precisely, if she has a particular kind of false belief about the action she is contemplating undertaking, then no amount of reflection can render that action autonomous. This suggests the need to include externalist conditions in an account of autonomy.

The autonomy-undermining potential of false beliefs has received surprisingly little philosophical consideration amid all the contemporary discussion of autonomy.2 Whilst it is occasionally mentioned in passing, rarely is it given sustained attention. This lack of

---

1 While this paper focuses on actions, the debates around autonomy extend to the conditions for autonomous values, desires, intentions and preferences. Though I will not be extending my conclusions here to cover these potential sites of autonomy, I am optimistic the general framework presented can be modified to accommodate these broader concerns. Whether such a framework can also be extended to considerations of more ‘global’ senses of autonomy is beyond the scope of this paper.

2 This is in contrast to the attention granted in medical ethics, via the notion of informed consent. See, e.g., (Savulescu and Momeyer 1997).
sustained attention would perhaps be less problematic if there were any kind of consensus around the relationship between autonomy and false beliefs, but in fact there is not. For instance, on the one side there are those such as Gerald Dworkin, who comments at the beginning of *The Theory and Practice of Autonomy* that we can grasp the difference between autonomy and liberty through reference to deception: deception undermines autonomy, but it does not undermine liberty (Dworkin 1988: 14). On the other side are those such as Nomy Arpaly, who by contrast declares that she doubts “whether anyone wishes to claim... that an *ill-informed* decision cannot be an instance of autonomous agency”, clarifying that ill-informed is intended to cover instances of deception (Arpaly 2005: 175). Clearly, there are divergent intuitions on the cost to autonomy of false beliefs.

Two exceptions to the paucity of critical attention given to false beliefs are a short discussion in Al Mele’s *Autonomous Agents* (1995: 179-82), and a recent contribution by Michael McKenna (2005). The arguments raised in these two conflicting accounts will provide the starting point for my claim that at least certain kinds of false beliefs are autonomy-undermining. In Section One, I outline the views of Mele and McKenna. In Section Two, I develop my argument for the autonomy-undermining potential of false beliefs. In Section Three, I introduce a schema for analysing which kinds of false belief compromise autonomy, and in what way.

1. **TWO VIEWS ON FALSE BELIEFS**

   **a) Mele’s View**

   Al Mele motivates his claim that false beliefs compromise autonomy through the use of intuition-pumping hypothetical scenarios. The most important of these involves the character of King George. Mele describes George’s predicament thus:

   [George’s] only access to the state of his kingdom is through his staff of advisors. George’s primary concern in life is to do what is best for his kingdom as a whole. The staff provide him with information, and he takes legislative and other measures on that basis. They also provide George with feedback about his measures, which information he takes into account in drafting further legislation and in modifying previous measures. They provide him, for example, with monthly figures on the gross national product, the distribution of wealth, the percentage of the population living below the poverty level, the king’s popularity, and the popular reaction to his legislative acts. However the staff have their own

---

3 Stoljar’s (2000) normatively substantive account of autonomy also includes quite considerable attention to false beliefs. However, her concern is with beliefs that are oppressive, over and above being false. Furthermore, the falsity enters her account through the norms that guide decision-making, rather than through the object of reflection. As such, her focus is slightly different to mine. I consider the relationship between Stoljar’s views and my own in Section Three
agenda, namely, their own enrichment at the expense of the populace. Knowing the king’s preferences, they systematically provide him with such false information as will lead him to make decisions that will further their aims (Mele 1995: 180-81).

Mele declares that George is clearly non-autonomous in this situation. This is because, in Mele’s words, George is “informationally cut off from ruling autonomously”. It is worth getting clear on how Mele takes this ‘information gap’ to undermine autonomy, since it will play an important preliminary role in my own account of false-beliefs.

According to Mele, being autonomous requires two distinct forms of control. On the one hand, we need control over determining our own values and ends. On the other hand, we need some degree of control over whether our acts succeed in realising those values and ends. George presumably has the first type of control: he has determined that he wants to rule so as to benefit the people, and there is no suggestion that this end is anything other than autonomously derived. No one has compelled him to value the end he holds; no one has manipulated him, hypnotised him, or brainwashed him; and there is no hint of any kind of evil demon so beloved of the philosophical imagination. However, George lacks the second type of control. Because of the false information he has received, George is unable to control whether or not his actions affect the achievement of his ends. The saturation of his informational context with false beliefs means that he cannot alter his behaviour to better achieve his ends, since he has no way of knowing that his actions are misfiring. Even worse, the harder he tries to achieve his ends the further away he will be from achieving them.

To summarise, according to Mele the deception faced by George undermines his autonomy because it blocks him from any kind of control over the extent to which his actions succeed in realising their ends. I believe that the conclusions Mele draws about George are correct. However, more needs to be said to validate these conclusions. I will take Mele’s reasons as a starting point for the argument that certain kinds of false belief undermine autonomy. First, though, we need to consider a counterargument to Mele’s conclusions.

b) McKenna’s View

Michael McKenna is one of the few theorists to explicitly defend the view that false beliefs do not undermine autonomy. He does so in the process of arguing that a distinction can be drawn between autonomy and moral responsibility, and that this distinction lies in the comparative stringency of two conditions: an epistemic condition, and a control condition (McKenna 2005). My focus here will be exclusively on his discussion of the epistemic condition, since it is here that he develops his view that false beliefs do not undermine autonomy.
McKenna argues that the epistemic condition for autonomy is weaker than the epistemic condition for moral responsibility. More precisely, he holds that if an agent attempts to do the right thing, but inadvertently does the wrong thing on the basis of a false belief, then she is not morally responsible. However, under identical circumstances she should be considered autonomous. Like Mele, McKenna motivates this claim through appeal to a hypothetical:

[T]he perfectly competent morally responsible agent Tal arrives at Daphne’s house and discovers her unconscious and in immediate need of the drug known as ‘The Good Stuff’. Urgently searching through Daphne’s cupboard, Tal finds a bottle marked ‘The Good Stuff’. He takes from it the prescribed dose and gives it to Daphne. Unfortunately Daphne’s fumbling pharmacist has accidentally given Daphne the wrong drug, ‘The Bad Stuff’, a drug that will kill people with Daphne’s condition. Daphne dies. Tal had no reason to be suspicious of Daphne’s pharmacist, and had very good reason to believe that the bottle in her medicine cupboard marked ‘The Good Stuff’ actually contained ‘The Good Stuff’ and not ‘The Bad Stuff’ (McKenna 2005: 208-09).

McKenna concludes from this parable that while Tal was not morally responsible (for he could not have known the consequences of his action, and he acted with good intentions), he is nonetheless autonomous. McKenna maintains that Tal autonomously gave the drug to Daphne.

McKenna’s justification for this conclusion rests on his claim that autonomy be understood in terms of acting on the basis of a principle that one endorses. In giving the drug to Daphne, McKenna points out, Tal was indeed acting on the basis of a principle that he endorsed: namely, always attempt to help those who are suffering innocently. “However we unpack the notion of ‘rule’ in ‘self-rule’”, McKenna argues, “there is no reason to think that Tal was not ruling himself in acting as he did” (2005: 209).

This now leaves us with two accounts that draw diametrically opposed conclusions about the relationship between false beliefs and autonomy. Moreover, these conclusions are drawn on the basis of hypotheticals that are structurally very similar.4 Where the reasoning

---

4 There is one disanalogy between the two hypotheticals that is worth mentioning. In the King George scenario, unlike in the case of Tal, George is deliberately misled in a way that makes him a conduit for others’ intentions. It may be supposed that this somehow renders George less autonomous than he would be were he to be labouring under identical misinformation that was not the result of human actions. As will become clear below, my theory does not make space for the source of the misinformation to affect the degree to which the agent’s autonomy is reduced. This may appear counter-intuitive. However, I suspect that the source of this intuition may have less to do with autonomy, and more to do with moral responsibility. Moral responsibility may push our intuitions on autonomy from two directions. First, when an individual is manipulated there is another agent who is morally responsible for her lack of autonomy. The graveness of the wrong of undermining another’s autonomy makes clear the harm befalling the manipulated individual, a harm that we may be less inclined to recognise in the non-manipulated but equally misled individual. It may simply be easier to identify a harm when we can identify the cause of that harm. This does not mean, however, that the
employed by Mele and McKenna comes apart is in how each construes the decision-making process necessary for autonomous action. For McKenna, that decision-making process has purely internal referents. Provided the agent endorses the desire that moves her to action, that action is autonomous. For Mele, by contrast, the decision-making process will at least in part be an external matter. While the desires that move the agent to action must meet a set of conditions to qualify as autonomous, as they do for McKenna, they must also meet the additional condition of connecting in the appropriate way with the action the agent takes herself to be considering. It is this difference that explains the divergent conclusions drawn by Mele and McKenna, and where our attention must focus in assessing the relevant merits of each view.

2. AUTONOMOUS ACTION

To determine whether false beliefs undermine autonomy, and if so in what way, we will need a clear account of precisely what autonomous action requires. It is only through understanding the mechanics of how autonomous action works that we will be in a position to see whether false beliefs disrupt that process. In part (a) of this section I offer an account of autonomous action. In a potentially question-begging manner, this account includes the requirement that the agent have competent comprehension of the action she is considering performing. I seek to justify this inclusion by showing how it is a necessary precondition of the kind of internalist reflection that virtually all autonomy theorists require for autonomous action. I will then draw on this account of autonomous action in part (b) to critique McKenna’s conclusions, before showing in part (c) how it can be used to support and augment Mele’s conclusions.

a) An Account of Autonomous Action

The account of autonomous action I present here takes as a starting point the idea that to act autonomously is to be engaged in the right kind of way with the decision to act. We can see that a condition of roughly this sort is required by considering some examples of paradigmatically non-autonomous action. I take it to be reasonably uncontroversial that I do not act autonomously if I act without even realising that I am acting. This might be the situation we find ourselves in, in those moments when, to our surprise and dismay, we deliberately harmed individual is any more harmed than the unintentionally harmed individual. Second, it is natural to assume that the manipulated individual is less morally responsible for what follows from her actions than the non-manipulated individual. This is particularly so if we feel that the non-manipulated individual is in some way culpable for the misinformation that guides her action. On my theory, the culpably misinformed individual is no more autonomous at the point of action than the manipulated individual, since all that matters is the way in which the false belief affects the decision-making process. However, this is compatible with the culpably misinformed agent retaining moral responsibility for her non-autonomous actions. Of course, this line of argument assumes that autonomy and moral responsibility can come apart, which is a position I cannot defend here. Thanks to an anonymous reviewer for this journal for raising this issue.
table into our mouth. While we might be able to point to all sorts of subconscious urges that explain our action, the fact that it occurs without us having decided to eat, or even wanting to eat, suggest that such actions are non-autonomous. Likewise, I take it to be reasonably uncontroversial that I do not act autonomously if I cannot prevent myself from acting, and I find myself effectively just a bystander to my actions.

From these intuitive beginnings, it is possible to start to build some necessary conditions for autonomous action. Paradigmatically autonomous action follows from a three-step decision-making process: first, comprehension of the potential action; second, reflection on whether to undertake the action; and third, resolution to undertake the action. In the wedges example above the first two steps are compromised, while in the bystander example the second two steps are compromised.

Importantly, while this three-part process describes paradigmatically autonomous action, it does not serve as the threshold for determining whether a given action is autonomous. I take it to be a reasonable desideratum of a theory of autonomy that it does not enjoin the conclusion that most people are non-autonomous most of the time. It a common complaint against some theories of autonomy that they are overly rationalistic, and invite the somewhat surprising conclusion that only philosophers are truly autonomous (c.f. Dworkin 1988: 17). Since much of our day-to-day activity doesn’t involve explicit deliberation, it would be helpful to develop a conception of autonomous action that allowed for, say, habitual activity to count as autonomous, provided this can be done in a way that is appropriately connected to the paradigmatic account offered above. To capture this idea, I suggest that what differentiates more autonomous from less autonomous action is not the fulfilment of the three step process of comprehension, reflection, and resolution, but the presence and accessibility of the necessary competencies to fulfil this process.

It will be useful to say a little more about what each of these competencies involves. As a starting point, we can say that in order to have autonomy over any given action A, the agent must be in a position to draw upon her competencies in these three areas.

**Comprehension**

In order to autonomously A, there are certain features of A that the agent must at that time be competent to comprehend. There are four distinct competencies at work

---

5 Diana Meyers (2005) has recently put forward a view that might seem to conflict with this scenario. Meyers describes a mountaineering accident in which she suffered multiple broken bones, and after which her body ‘took control’ and facilitated her descent down the mountain. Meyers argues that this was not an instance of non-autonomy, but rather an instance of autonomy skills exhibited by the self-as-embodied. I see the view I develop here as compatible with this claim. Not all automatic action need be considered non-autonomous, as I explain below. As will become clear, Meyers’ actions in fact reflect just the competencies that I claim are necessary for autonomous action.

6 Some philosophers describe hypnosis or addiction in these terms. I doubt whether this is in fact an accurate description of such experiences (see, e.g., (Levy 2006)), but little hinges on the matter here.

7 Some exceptions to these conditions will be considered in Section 3 (b) below.
here. First, the agent must be competent to recognize what action A is. Second, she must be competent to understand what plausibly follows from doing or refraining from A. Third, she must be in a position to recognize that A is practically available to her. Fourth, and more tentatively, she must be in a position to recognize that ~A is practically available to her.

**Reflection**

In order to autonomously A, the agent must be competent to reflect on certain features of A. This will primarily involve the competency to weigh the pros and cons of doing A before committing to action, which in turn involves competent reflection on the quality of one’s own desires for and valuing of A and its consequences.

**Resolution**

Finally, in order to have autonomy over A the agent must have the competency to follow through on her reflection by forming an intention and performing it. If an agent reflectively endorses A, but then simply happens to drift into it despite not having the competence to form a corresponding intention, then she does not autonomously A. Similarly, if she forms an intention but then performs an unrelated act, she has not acted autonomously.

What determines whether an agent’s action is autonomous, then, is whether the competencies within these three areas are accessible at the time of action. Admittedly, if the agent has not in fact exercised a given capacity, it may be extremely difficult in practice to ascertain whether or not it was accessible to her. This does not mean, however, that the distinction is meaningless. The distinction between exercise and accessibility of a competency can best be illustrated through analogy. Being in control of a car requires certain mental and physical competencies. However, one can be in control of a car without those competencies being exercised. For instance, when driving a familiar route it is common to find oneself at the destination without consciously having gotten oneself there. We might think of this as a kind of subconscious exercise of one’s mental and physical competencies. Additionally, one may have employed cruise control on route. In this case, some of the competencies were not even being subconsciously exercised. They were, however, accessible. At any time, the driver could have ‘re-engaged’. The utilisation of cruise control, and hence the failure to exercise one’s competencies, does not undermine one’s control over the car. Similarly, habitual action can still be autonomous, provided the

---

8 More will be said to qualify this condition later, since as it stands it potentially has the unfortunate effect of requiring something approaching omniscience in order for action to be autonomous.

9 This description of the reflection competencies is kept deliberately loose here, since my goal is to show that however the details of reflection are construed, autonomous action requires comprehension as well. If successful, this would mean that for any current account of autonomous action, there is an additional condition that has yet to be acknowledged.
agent is in a position to re-engage her competencies – provided, that is, she remains in the driver’s seat.

For the purposes of this paper, the first competency within the category of comprehension is going to do most of the heavy lifting.\textsuperscript{10} It is necessary, then, to defend its inclusion in my account of autonomous action. Why should we think that I must be competent to comprehend the act I am considering undertaking, for that act to be autonomous? There are two ways in which the inclusion of this condition can be justified. First, it can be shown that the reasons for including it are parallel to the reasons for including a minimal rationality condition for autonomy. If one were to accept a rationality condition, therefore, one should also accept comprehension conditions. More significantly, I will argue that a comprehension condition is necessary to make sense of internal reflection. For any account of autonomy that privileges the quality of an agent’s reflection (and there are not many that don’t), there is good reason to supplement it with a comprehension condition.

To start with rationality: Let’s say that I want to achieve condition B. I know that A leads to ~B; but I do A with the intention of bringing about B.\textsuperscript{11} Have I autonomously A-ed? Those who support a minimal rationality condition for autonomy, such as John Christman (1991: 13-16), will say no. The best reason for saying no, I want to suggest, has to do with what it means to be self-governing over actions. Autonomous action involves, at the very least, acting on the basis of an intention. But to act on the basis of an intention requires there to be some connection between my action and my intention. I act \textit{in order to} fulfil my intention. As such, there needs to be some degree of coherence between my intention and my action if that action is to be considered self-governed. The problem with manifest irrationality of the type described above is that it disrupts this coherence. My intention and my action are irrevocably separated, since my action actually leads away from the realisation of my intention – which, it should be stressed, was the reason I had for acting in the first place.

Now while failure of comprehension is not the same thing as failure of rationality, its effect on this necessary correspondence between intention and action \textit{is} the same. Just as irrationality breaks the connection between intention and action, so too does failure to comprehend what it is I am about to do. This can perhaps best be illustrated through analogy with hallucination. Let’s say I form the intention to eat a piece of cake. To fulfill this intention, I go to what I take to be the fridge and retrieve what I take to be a delicious cream-filled Swiss roll. In fact, what I have done is gone to the cupboard and retrieved a roll

\textsuperscript{10} I focus here on the first condition, since this will be sufficient to show that McKenna is mistaken in his conclusions about false beliefs. The relevance of the second condition, that the agent comprehends what necessarily follows from an action, will be drawn out in Section 3.

\textsuperscript{11} This is distinctly different from a situation in which I know that A leads to ~B but I do it anyway, in the absence of an expectation that it will bring about B. Such action may be called irrational (or perverse, or weak-willed), but it is not my concern here. On the account I put forward, such action will be considered autonomous, provided it meets the necessary conditions.
of dog food, which I am eagerly bearing towards my mouth. Consider the parallels with rationality. If I had knowingly retrieved the dog food with the intention of satisfying my desire for cake, it would have been irrational, and hence non-autonomous: there is no way the action could have satisfied my intention, and so the action does not reflect my self-governing command to eat cake. Precisely the same problem occurs when I unknowingly retrieve the dog food. It cannot satisfy my intention, and so the action is divorced from my self-governance. The order I give myself is to satisfy my desire for cake, yet that is not possible given the action I perform. As such, the action was not under the direction of my will.

This conclusion can be further buttressed through consideration of the connection between comprehension and reflection. As I noted above, much attention has been paid to determining the conditions that must obtain over an agent’s reflection, for her actions to be considered autonomous. However we understand these reflective conditions: whether it requires that we have an appropriate second order desire towards the desire that moves us to action (Frankfurt 1971); or that we establish coherence between the preferences that move us to action and our conception of the good (Ekstrom 1993); or that we not feel alienated by a desire or value in light of its history (Christman 1991); what these theories have in common is an exclusive focus on the quality of the agent’s motivations for action. By focusing on the motivation for action, and not the action itself, such theories are subject to a considerable blind spot: No matter how autonomous an agent’s motivations are, the action itself cannot be autonomous unless it is appropriately connected to the motivation behind it.

Refocusing attention on action rather than intention is crucial. The reason why we should be concerned with higher order desires, or the importance of valuing and planning, or non-alienation, is that they seek to address the requirement that self-governance issues from the self. But to leave the issue here overlooks the requirement that self-governance involves governance. While it is crucial that the actions we perform follow from motivations that are in some sense our own, it is equally crucial that the actions we perform are in fact those actions that we intended to perform. In acting autonomously, we are expressing our motivational states through our action. We do not just reflect, or plan, or value in the abstract. We reflect, or plan, or value because we have to decide what to do (Bratman 2000). There are thus two ways in which action can fail to be autonomous. It can fail, as many have noted, because the motivation from which it originates is in some sense not our own. But equally, it can fail because the action is divorced from our motivation. Such a divorce will occur when we do not have the competency to comprehend what it is we are doing when we act. This can come about through hallucination, as with the Swiss roll case, but equally it can come about through more prosaic false beliefs. Reflection alone is insufficient to render action autonomous, since the conclusion of our reflection needs to be expressed in that action.
There is, moreover, a further reason to incorporate comprehension into an account of autonomous action. Not only must the command that I issue to myself be followed by action that corresponds to it, as noted above, but the command itself must reflect my will. This is potentially compromised if the object of my reflection is not what I take it to be. If I am mistaken about what the potential action is that I am considering performing, or mistaken about what would follow from performing it, then it is difficult to see how my reflection about what to do can express my will (c.f. Kawall 2010: 124-25). We can see this by re-examining the cake/dog food scenario from the perspective of Frankfurt’s second order desire model (though the problem will generalise to all accounts focused on the quality of one’s motivation). Let’s imagine that I have a desire to eat cake. For the eating of cake to be autonomous, on Frankfurt’s model, I would have to desire to desire to eat cake. What might explain such a second-order desire? Perhaps I reflect on the enjoyment that cake eating will bring to me. Or perhaps, selflessly, I consider how eating the cake would aid in my flatmate’s diet. Whatever reason explains the second-order desire I have, if that reason makes reference to the cake actually being cake as opposed to dog food (as the two suggested here most certainly do), then the only action that could autonomously follow from it would be eating cake. I may have an autonomous desire to eat cake, but that does not mean that anything I subsequently put in my mouth is put there autonomously.

b) Why McKenna Is Wrong About False Beliefs

Armed with this account of autonomous action, we are in a position to counter McKenna’s claim that false beliefs do not compromise autonomous action. McKenna argues that Tal was autonomous, because when Tal administered the drug in the vial he was acting in accordance with his principles. However, at least on one reading of what occurred in Daphne’s house, administering the drug in the vial did not accord with Tal’s endorsed principle. Since administering the drug in the vial was inevitably going to be fatal to Daphne, given what was in fact in the vial, Tal’s action was incompatible with his principle of helping those suffering innocently.

There are two ways in which we could make sense of McKenna’s position here, though as we will see neither of these is successful. First, we could read McKenna as offering an exclusively internalist account of autonomous action. On the second reading, we could read him as allowing a very limited externalist aspect to autonomous action.

On the first reading, determining the autonomy of an action begins and ends with the command given to oneself to act. McKenna at times explicitly supports this position: “[autonomy’s] scope comes to an end at the moment in which an agent performs simple mental actions, such as deciding or choosing” (McKenna 2005: 229, fn 14) As such, if the command to act accords with one’s self-governing principle, the act which follows is autonomous. If this is the correct interpretation, it means that for McKenna the content of the act itself is irrelevant to the determination of whether the agent acted autonomously.
Taking the question of whether or not Tal acted autonomously from this perspective would thus turn on whether what we mean by autonomous action refers exclusively to the decision to act, or incorporates the action itself. Considered from the perspective of the autonomous decision-making procedure outlined above, it is clear why the action itself must be incorporated into our determination of autonomy. While I agree with McKenna that Tal’s reflective competencies are unadulterated, and as such his intention to give the contents of the vial to Daphne is autonomous, he fails to be self-governed unless the action that he performs is appropriately connected to his reflections. McKenna’s conclusions about Tal, and subsequently his claim that false beliefs are compatible with autonomous action, rely on denying this connection between reflection and comprehension.

To fend off any lingering attachment to this purely internalist account of autonomy, a short example should help. Imagine that Betty commands herself to give money to a beggar, in accordance with her endorsed principle of helping the poor. According to McKenna, her action is rendered autonomous at the moment of making this command. Unbeknownst to Betty, however, she has issued the command whilst half asleep. Instead of proffering a few dollars to a homeless woman, she in fact whacks her sleeping partner in the face with her enthusiastically outstretched hand. *Pace* McKenna, it seems very odd to say that this action was autonomous. While the command itself was at least arguably autonomous, the action Betty actually performed – hitting her partner in the face – was most certainly not.

Reading McKenna as purely internalist commits him to the counter-intuitive notion that the success conditions for autonomous action bear no relation to the action performed. While McKenna appears to invite this reading, it is worth considering whether his account can be interpreted more plausibly as including an implicit, though weak, externalist criterion. On this reading, there is something about the action performed by Tal that sufficiently maps onto the command to act to render that action autonomous. For instance, Tal’s order to himself presumably contained such sub-instructions as ‘open the vial’ and ‘insert the contents of the vial into Daphne’s mouth’. These actions were in fact performed by Tal, in a way that Betty’s offer of spare change to the beggar was not performed. Some kind of correspondence condition between the command and the action would potentially allow us to differentiate the two cases, making it coherent to claim that Tal acted autonomously whilst denying that Betty did.

To see how this pathway is closed to McKenna, we need to consider the issue of act individuation.12 Part of how McKenna explains his claim that Tal acted autonomously is to note that Tal gave the drug in the vial to Daphne, which is precisely what he intended to do. For that claim to be convincing, however, we would need to collapse the events that followed from Tal’s decision into a single act. In other words, we would need to see no separation between:

---

12 Thanks to Al Mele for helping to clarify some of the following issues.
a) Tal autonomously tried to save Daphne

b) Tal autonomously gave a drug to Daphne

c) Tal autonomously gave ‘the bad stuff’ to Daphne.

By claiming generally that Tal acted autonomously, McKenna does not allow for the fact that there may have been elements of the chain of events that were performed autonomously, and others that weren’t. He doesn’t allow for any act individuation at all. Crucially, this means he doesn’t allow for the fact that precisely the element that was not performed autonomously was the act of giving ‘the bad stuff’ to Daphne.

The reason why giving ‘the bad stuff’ to Daphne was non-autonomous, while other actions (i.e. trying to save Daphne) plausibly were, rests on the connection between intention and action outlined above. For an act to be autonomous, it needs to be connected in the right way with the intention to act. It needs to be the same act that the agent has reflected on and resolved to perform; otherwise we are back to the cake/dog food scenario from above. While it is plausible to say that Tal reflected on and resolved to perform actions a) and b), it is not at all plausible to say the same for action c). Given that the drug in question was in fact ‘the bad stuff’, it is thus implausible to say that Tal gave autonomously gave the drug to Daphne (though of course there were other, related, actions that he did perform autonomously).

We have seen that McKenna’s claim that Tal acted autonomously in giving the drug to Daphne does not survive two possible readings of his argument. On the first reading, McKenna would be forced to concede that the success conditions for autonomously performing an action have no connection to whether or not the action is actually performed. Unless we want to say that there are only autonomous intentions, and no autonomous actions, this path leads to a dead end. Alternatively, on the second reading, McKenna could be claiming that there is sufficient correlation between Tal’s intention and the act that he in fact performed. This would only lead to the conclusion that Tal acted autonomously, however, if we do not recognise any individuation in the potential descriptors of Tal’s action. While there are certain actions Tal autonomously performed following his decision, giving the drug to Daphne was not plausibly one of them.

All of this might raise a very distinct concern. Specifically, it might raise the concern that my account requires an action to be successful if it is to be autonomous. This is precisely the reason why McKenna objects to Mele’s conclusions regarding false beliefs, so it will be a useful entry point to our examination and defence of Mele’s position.

c) Why Mele is Right About False Beliefs
As noted in Section 1 (a) above, Mele holds that for an action to be autonomous the agent must have some degree of control over whether it achieves its goal. McKenna objects to this condition on the basis that success is an inappropriate condition for autonomy; the effectiveness or otherwise of my action has no bearing on whether that action is autonomous. This seems right. We should certainly leave open the possibility that someone can be self-governing and unsuccessful. For instance, let’s imagine that Jill whacks Jack over the head with a crowbar, with the intention of killing him. If Jack miraculously survives, it certainly doesn’t mean that Jill did not act autonomously. Does this mean that Mele is wrong, and the control condition must be abandoned?

To see why it does not, we need to look more closely at the detail of Mele’s control condition. As Mele articulates it:

>[A] sufficient condition of S’s being informationally cut-off from autonomous action in a domain in which S has intrinsic pro-attitudes is that S has no control over the success of his efforts to achieve his end in that domain, owing to his informational condition (Mele 1995: 181, emphasis added).

Mele’s condition is more subtle than McKenna takes it to be. Rather than claiming that actions need to be successful in order to be autonomous, Mele is claiming that an agent’s autonomy is undermined if, due to the quality of her information, there is no connection between her intention and the action she takes to realise it.

In the Jill and Jack example Mele’s control condition is met, so that despite Jill’s failure to achieve her end she acted autonomously. This is because there is an appropriate connection between Jill’s intention and the action she performs to realise it. She is not ‘informationally cut-off’ from achieving her goal, since there is nothing in the way the world is that would suggest whacking Jack over the head with a crowbar wouldn’t be a good way of achieving the goal of killing him. Moreover, and in stark contrast to King George’s predicament, there is every reason to think that if Jill tried again with a little more vigour she would achieve success.

Success is not what Mele’s control condition requires. Rather, his condition requires that the possibility of success is not thwarted due to informational gaps. Mele is including an explicitly external component to his account of autonomy. An agent’s decision-making process must map onto the world in some way, and the agent’s autonomy is undermined to the degree that it fails to do so.

Mele’s control condition is supported by the account of autonomy developed in this paper. Mele’s claim that autonomy is undermined through failure to have control over the success of one’s actions, where this failure is due to informational constraints, appears to be drawing on the same reasoning as my claim that autonomy is undermined through failures of comprehension. False beliefs create a disconnect between one’s actions and the
reflective decision to act. While Mele’s account draws attention to this disconnect, my account has provided an explanation for why this disconnect is autonomy-undermining.

3. VARIETIES OF FALSE BELIEF

I have argued that false beliefs undermine autonomy, because they compromise the decision-making process necessary for autonomous. In this final section, I clarify and elaborate on this claim. Doing so will involve identifying the kinds of false beliefs that are relevant to autonomy. This will involve not only explaining more fully how false beliefs about the act in question undermine autonomy, but also exploring the possibility that false beliefs about what follows from an action, and false beliefs about the norms motivating action, likewise undermine autonomy. I will first lay out a general classification of false beliefs and how they affect autonomy, before modifying the account to accommodate potential objections.

a) Classifying false beliefs

The different ways in which false beliefs compromise autonomy can be seen in the following example of ‘drinking the Kool-aid’. In this example, an agent drinks the contents of a glass containing cyanide. In each version, a different false belief affects the agent’s decision-making procedure.

**Version 1**: Carla drinks the contents of the glass in the belief that it’s actually Kool-aid. She has the mistaken belief that the glass does not contain cyanide.

**Version 2**: Carla drinks the contents of the glass, fully aware that it contains cyanide, and understanding what cyanide is. However, she believes that drinking the cyanide will not kill her, but will rather transport her to an alien planet.

**Version 3**: Carla drinks the contents of the glass, fully aware that it contains cyanide, and fully aware that it will kill her. She drinks it because she believes that she does not deserve to live, after having brought shame on her family through an extramarital affair.

These three versions reflect three different ways in which a false belief may affect an agent’s decision-making procedure. Since each interacts with a different step in that decision-making procedure, they can be classified according to the particular competency they undermine. In Version 1, we have a scenario very similar to that faced by Tal, in McKenna’s hypothetical. Carla’s false belief results in a failure of her competency to recognise the act that is being performed. I call this a primary false belief.

---

13 It is crucial here that the agent actually understands that cyanide contains properties that are poisonous, and usually kill people. In the absence of this understanding, the second version collapses back into the first, becoming a situation of not comprehending what is in the glass.
Failure of competency to recognise the act that is being performed initiates a chain-reaction of competency failure throughout the decision-making procedure. Without recognising the act in question, Carla is not competent to determine what would likely follow from performing it. Without understanding what would follow from an act, she is not competent to reflect on its desirability or coherence with her plans and values.

In Version 2, the false belief enters the decision-making process further along the chain. Here, Carla is competent to comprehend the act she is performing. This false belief results instead in a failure of competency to recognise what follows from the act. I call this a secondary false belief.

Version 2 is analogous to King George’s situation. He, too, perceived the actions he was performing, but did not have the competence to understand what followed from them. As with a primary false belief, a secondary false belief results in a disconnection between the agent’s actions and the intentions that drive them. Once more, there is a chain reaction along the decision-making procedure. Since Carla does not understand what follows from her action, she cannot reflect on the desirability of that action, or its coherence with her values or plans. While a primary false belief interrupts this chain at an earlier stage, a secondary false belief is equally autonomy-undermining. What matters for an action being autonomous is the extent to which the decision-making process as a whole is compromised. Any disruption at the level of perception will upset this process, and cut off the agent’s ability to competently reflect and resolve to act.

This brings us to Version Three. In this scenario, Carla comprehends the act she is performing; she understands what will follow from that act; she has reflected on the desirability of the act; and she has resolved to act. Thus far, it seems to be a paradigmatically autonomous action. There is, however, a false belief at work: Darla has a false belief about her own moral worth, and it is this false belief that guides her action. On these grounds, her action would presumably not be considered autonomous by Natalie Stoljar (2000). Since the false belief enters into the quality of the agent’s reflection, I call this a tertiary false belief.

While I am happy to agree with the determination that Carla’s actions may not be autonomous in this scenario, I do not agree that the reason for this can be traced to the falsity of her beliefs. This is because, unlike competence in perception, competence in reflection does not require reflection to track objective values or reasons. What matters is that reflection tracks values or reasons that are the agent’s own. In this scenario, there is no reason to believe that Carla is mistaken about what she wants to do.14

14 I leave aside the related problem of being mistaken about what one actually values. My concern is with the way in which accounts of autonomy have tended to overlook the importance of external criteria. Since false beliefs about one’s own motivation are already a consideration for internalist accounts, I will not discuss them further here.
It has been argued by a number of theorists that lack of self-worth or self-respect is inimical to autonomy (Benson 1994; Dillon 1997; Govier 1993; Hill 1991; McLeod 2002). Since Carla’s actions reflect a significant lack of self-worth, we might claim on those grounds that she is non-autonomous. However, the reason why these theorists claim that lack of self-worth or self-respect compromises autonomy is not because such attitudes reflect false beliefs. For those who posit a self-worth or self-trust condition, it is because such attitudes are themselves causally or conceptually necessary for autonomous action. For Benson, for instance, autonomy is understood in terms of having a sense of oneself as worthy to answer for one’s actions, a stance that requires a sufficient degree of self-worth. According to Govier and McLeod, meanwhile, the competencies that are necessary for autonomous action, such as being able to rely on one’s own critical reflection, are dependent on an attitude of self-trust (Govier 1993; McLeod 2002).

This is not the place to enter into a detailed examination of the plausibility of such claims. Suffice to say that if such accounts are correct, certain normative beliefs will be incompatible with autonomous action. However, this will not be because such beliefs are false, since the compromising effect of low self-worth would remain even if the belief were in fact true.

b) Modifying the theory

What I have presented above is a basic schema for understanding how false beliefs compromise autonomy. Primary false beliefs relate to the nature of the act that I am contemplating undertaking; secondary false beliefs relate to the consequences of the act I am contemplating undertaking; and tertiary false beliefs relate to the norms that govern my decision-making. I have suggested that tertiary false beliefs do not necessarily compromise autonomy. Certain normative beliefs may have a deleterious effect on autonomy, but this is due to their direct incompatibility with competent reflection, rather than with their falseness per se. It must now also be considered whether and under what conditions primary and secondary false beliefs do not impede autonomy.

False beliefs undermine autonomy insofar as they compromise the relationship between intentions and actions. However, some false beliefs will simply be irrelevant to that process. For instance, if the agent in our Kool-aid example falsely believed that the glass contained strychnine rather than cyanide, her autonomy would most probably not be undermined. This is because the reasons for drinking or not drinking strychnine will predominantly map onto the reasons for drinking or not drinking cyanide. She would therefore be competent to

15 I do in fact hold that they are highly plausible. On my account, autonomy requires the accessibility of a range of competencies. These competencies appear vulnerable to certain self-regarding attitudes, such as self-distrust, in just the way Govier notes. I take it, however, that the link between self-regarding attitudes and the autonomy competencies is an empirical rather than conceptual one, and thus proving the connection would require psychological evidence.
understand the consequences of drinking the contents of the glass despite having a false belief about it, and correspondingly competent to reflect on whether those consequences are desirable. Similarly, we could tweak the Tal example so that he falsely believes the vial contains The Good Stuff, when in fact it contains The Excellent Stuff. Again, the reasons Tal has for administering The Good Stuff would carry over to delivering The Excellent Stuff. In doing so, his decision-making procedure would not be compromised. To put it in McKenna’s terms, he would still be acting on the basis of a principle he endorses.

There may also be instances in which false beliefs not only fail to undermine autonomy, but instead positively augment it. Autonomy requires that I take action A to be practically available to me. Sometimes, a false belief will bring about a belief that A is practically available to me, whereas a true belief would cause me to believe the action was closed off to me. For instance, if I am confronted with a crevasse as I attempt to escape an encroaching bushfire, whether the option of leaping across is open to me will depend in part on my beliefs about how wide the crevasse is. If I believe truly that it is three metres wide, I may be paralysed with fear, and thus unable to act. If I believe falsely that it is only two metres wide, this action may become available to me.16 A false belief here opens up scope for autonomous action that would be closed off if I had a true belief.

Autonomy also requires that I take ~A to be practically available to me. I do not act autonomously if I act on the basis of a belief that I could not have done otherwise. As Frankfurt (1969) has convincingly argued in the context of moral responsibility, however, it is far less clear that alternative action must in fact be practically available to the agent. Considering this problem from the perspective of false beliefs, we have reason to agree with Frankfurt. If I falsely believe that ~A is available to me, the relationship between my intentions and my actions remains intact. This would not be the case in the inverse scenario, namely if ~A were in fact open to me, but I falsely believed that it wasn’t. In this scenario, my intentions are potentially corrupted. It is harder (though perhaps not impossible) to gauge one’s desire for A if A is all that is believed to be available, as can be seen with the problem of adaptive preferences (Elster 1985; Nussbaum 2000).

As we have seen, false beliefs may under certain conditions fail to undermine autonomy. The next issue to consider is whether those false beliefs that do undermine autonomy all do so to the same degree. The first thing to note here is that autonomy itself admits of degrees. While it may on occasion be necessary to invoke a threshold for autonomy (i.e. when setting guidelines for paternalistic interventions) such thresholds reflect social and political needs for practical determinacy, rather than identifying a meaningful tipping point in individual autonomy.17 As Meyers (1989) observed more than two decades ago, once

16 This is deliberately different from cases where believing something makes it the case, as in William James’ (2010) willing to believe.

17 A good analogy for this is the voting age restrictions most states employ. The right to vote is taken to be contingent on a certain level of rationality, autonomy and social understanding. No one would suggest, however, that there is a meaningful shift in these competencies at the moment the individual turns 18.
autonomy is recognised as the exercise of a range of competencies it follows that autonomy can be exercised more or less well, in accordance with the development and deployment of the relevant competencies.

This has implications for the relationship between autonomy and false beliefs. We have seen that false beliefs affect autonomy by compromising the competency to comprehend a potential action, both in terms of what that action is, and what will follow from that action. The compromising of comprehension in turn affects reflection, as the agent is potentially unable to connect her reflection to the action when that action is misunderstood. As should now be obvious, though, reflection can be compromised to a greater or lesser degree, depending on the content of the false belief. There are some false beliefs that will not affect reflection at all, as we have seen with the cyanide vs. strychnine example, since the reflection process will be tracking the same consequences, and therefore the same values and preferences within the agent. In addition, there are some false beliefs that will only partially affect reflection.

We can see how this works by returning to the example of Tal. In McKenna’s rendition, Tal believes that the drug is 100% effective, whereas in fact it is 100% fatal. If we now alter the scenario, so that Tal still believes that the drug is 100% effective, but in fact it is 30% fatal, would we still want to say that Tal acted non-autonomously? Much will depend here on Tal’s personal reasoning process, and consequently on his values and preferences. It may be that Tal is highly risk-averse, such that he would never dream of administering a drug with such a high percentage risk of mortality. In that case, Tal’s action in administering the drug does not truly flow from his self. However, it may be that Tal is a risk-taker, and the knowledge that the drug has a 30% chance of mortality would barely register in his reasoning process. In that case, we would want to say that Tal acted autonomously, since the false information has not created a disconnect between his reasons for action and the action he performs. Within this range, however, there are many other possibilities. The more the true knowledge of the drug’s potential mortality would alter Tal’s reflection, the more the false belief compromises his autonomy. And conversely, the less the true belief would alter Tal’s reflection, the more autonomous the resulting action would be.

This clarification also helps alleviate a potential problem facing the second comprehension condition – namely, the condition that the agent must understand what plausibly follows from an act. As it stands, that condition may raise the worry that the agent would only be autonomous if she were omniscient. With the aid of our scalar conception of autonomy, however, this worry dissipates. It may well be that the most autonomous agent would be an omniscient agent, but it does not follow that lack of omniscience fully undermines autonomy. Once more, what matters is the degree to which the false belief affects the agent’s reflective process.

A final observation is in order. McKenna holds the intuition that not only is Tal autonomous, but King George is too. Part of what seems to be driving this intuition for McKenna is the
fact that both Tal and George succeed in acting in the world, in at least a recognisably similar manner to the way they intended to act. Given the acknowledgement that autonomy comes in degrees, I am happy to concede that both Tal and King George do exercise a very minimal degree of autonomy. This returns us to the issue of act individuation. We can see why their stock of autonomy is not entirely empty by comparing their cases to one in which the agent intends to act in a certain way but instead performs an entirely different act. For instance, if instead of administering the drug as intended Tal had inadvertently found himself circling Daphne while clucking like a chicken, his autonomy would no doubt have been much more seriously compromised. Control over our bodies is an important sphere of autonomy, and insofar as that remains we can truthfully say that however false the beliefs that underpins an agent’s actions, she is at least to some degree acting autonomously when her body responds to her intentions.

To conclude, determining the degree to which false beliefs undermine autonomous action relies on two distinct points of inquiry. First, it is vital to be clear on precisely which act we are interested in making the determination about. Much confusion follows from failing to do so, as can be seen in McKenna’s discussion of Tal. Acts will often be nested, which allows us to make nuanced determinations regarding the point at which autonomy starts to fail, provided we have individuated the range of acts under question. Second, we need to investigate the connection between the false belief and the agent’s reflection on the act/s in question. To the extent that the false belief figures in the agent’s reasoning, in terms of the likelihood that had the false belief not present the agent would have acted otherwise, then to that extent her autonomy is compromised. Act individuation and influence on reasoning are thus deeply interconnected, since how we individuate the act will determine whether we think the agent would have done otherwise. To close with an example, we can return to Tal. If the question ‘did Tal act autonomously’ refers merely to the physical movement of placing the vial in Daphne’s mouth, then Tal acted autonomously. He would have performed just the same movement if the drug had in fact been the Good Stuff. On the other hand, if we are interested in whether the giving of the Bad Stuff was done autonomously, then the answer depends on how knowledge of the drug’s properties would have affected Tal’s reasoning. This depends in turn on the likelihood of the Bad Stuff having fatal effects, combined with Tal’s risk-averseness. The more Tal’s intentions become divorced from the action he ultimately performs, the less that action is autonomous.


