

An elusive challenge to the authorship account: commentary on Lawlor's "Elusive reasons"

LUCA FERRERO

ABSTRACT *Lawlor argues that social psychological studies present a challenge to the authorship account of first-person authority. Taking the deliberative stance does not guarantee that self-ascriptions are authoritative, for self-ascriptions might be based on elusive reasons and thus lack agential authority (i.e. they are no guide to the subject's future conduct). I argue that Lawlor's challenge is not successful. I claim that we can make sense of the nature and importance of agential authority only within the framework of the authorship account. Agential authority is part of the regulative ideal of the deliberative stance, but its lack does not undermine the first-person authority of self-ascriptions, since first-person authority is primarily a matter of deliberative authorship.*

1. Introduction

In "Elusive reasons: a problem for first-person authority," Krista Lawlor presents a challenge to the authorship account of first-person authority. According to this account, whose most sustained defense is to be found in the work of Moran (2001), the distinctive first-person authority of self-ascriptions of attitudes is due to one's *authoring* these attitudes by taking deliberative responsibility for them. That is, by making the attitudes responsive to one's sense of the best reasons in their support; a responsiveness secured by a readiness to take up the deliberative question about the existence and nature of these reasons.

According to Lawlor, the authorship account (*AA* hereafter) does not cohere with the commonsense view of first-person authority since it takes as authoritative self-ascriptions that appear to be "hollow." These are the self-ascriptions that, unbeknownst to the subject, are based on what Lawlor (2003, Section 8) calls "elusive reasons," i.e., on biased and manipulated considerations.

Self-ascriptions based on elusive reasons lack what Lawlor calls "agential authority." We sense that these self-ascriptions are, as she says, "no guide to how the person will ultimately choose, act, or feel about her choices and actions in the future" (Lawlor, 2003, Section 9). Take the case of a subject *S* who sincerely

Luca Ferrero, Department of Philosophy, University of Wisconsin-Milwaukee, PO Box 413, Milwaukee, WI 53201, USA, email: ferrero@uwm.edu

claims that she believes that p and does so by taking deliberative responsibility for this self-ascription. Let's assume that, unbeknownst to S , the self-ascription is based on biased and manipulated reasons. If so, we cannot rely on S 's self-ascription "I believe that p " in predicting what S is going to do in the future. As far as we can tell, S might or might not believe that p when the time of action comes, and thus we cannot predict her future conduct to the extent that this conduct is affected by S 's future believing that p .

In this paper, I will argue that Lawlor has not presented a real challenge to AA . I will not deny either the phenomenon of elusive reasons or the importance of agential authority. My contention is that we can make sense of the nature and importance of agential authority only within the framework provided by AA . Lawlor is correct in calling attention to the importance of agential authority in accounting for first-person authority, but she is mistaken in suggesting that AA cannot properly account for this authority. Elusive reasons therefore present no real challenge to AA . Or so I will argue.

2. First-person authority

When considering the phenomenon of first-person authority, we must distinguish between different senses in which self-ascriptions can be authoritative. The standard sense in which S 's self-ascription can be said to be *authoritative* is that it correctly indicates S 's attitude. In this sense, saying that S speaks with authority in making the self-ascription means that she is knowledgeable about her own attitudes. S is knowledgeable about these attitudes in a way that is not just that of someone who is an expert at this subject matter. First-person authority is distinctive in that S is in a privileged position with respect to the self-ascribed attitudes, a position that no one else could occupy. The special position might not guarantee that the self-ascriptions are infallible, but S has at least a good *prima facie* claim to the truth of the self-ascription (Moran, 2001, p. 10). Moreover, S 's self-ascriptions reflect the immediacy and transparency of S 's relation to her own attitudes.

A traditional way of accounting for these features is to take S to be a privileged expert about her mental goings-on in that she has an exclusive evidential access (usually in the form of introspection) to her attitudes. First-person authority is here a matter of the authority of expertise about one's psychology. A different kind of account makes this authority a matter of social concession, so that S 's self-ascriptions are authoritative because they are accepted as such by convention. Each subject has the authority of an absolute sovereign—so to say—about one's attitudes, so that one's self-ascriptions settle what attitudes one has as the proclamations of the sovereign settle the law of the land [1].

AA rejects these solutions. According to AA , self-ascriptions are authoritative because the subject *authors* them, not by decree, but by taking a *deliberative stance* toward the reasons that justify the self-ascribed attitudes. According to AA , as Lawlor (2003, Section 3) correctly puts it,

[O]ne takes responsibility for one's attitudes by being ready and disposed

to take up deliberative questions about what one's attitudes are to be in light of one's reasons. And that is what it takes to speak one's mind with special, first-person, authority.

The subject has what might be called the "authority of authorship." She is the only one who can acquire the attitude by making up her mind about the attitude [2].

Notice that according to *AA* *S*'s self-ascription of an attitude depends on *S*'s recognizing and endorsing the reasons that make the self-ascribed attitude justified. In the case of a belief that *p*, for instance, *S* ascribes it to herself by considering whether *p* is the case. She answers the question about her mental state by an outward not an inward glance (Moran, 2001, pp. 60–62, 134). In the first-person case the question whether I believe that *p* collapses into the question whether *p* [3]. The possibility of false beliefs, however, shows that there is still a distinction between the self-ascription of the belief and the self-ascribed belief. If *p* is not the case, *S*'s belief that *p* is false. In this case, *S* does not speak with authority with respect to the question whether *p*. But to the extent that she does not realize that *p* is false, *S*'s speaks with authority with regard to her *believing* that *p*. The authority of authorship guarantees that *S* is knowledgeable about her attitudes, not that she is knowledgeable about the correctness of her attitudes (see Moran, 2001, p. 74).

In objecting to *AA*, Lawlor (2003, Section 8) claims that there are circumstances in which, even if *S* sincerely takes deliberative responsibility for her attitude (say, the belief that *p*), the self-ascription sounds hollow and empty because we can legitimately wonder whether this is *really S's* attitude (whether *p* is really what *S* believes). To show this, Lawlor uses the results of the experiments of both Seligman *et al.* (1980) and Wilson *et al.* (1995). These experiments show that a subject can be induced to self-ascribe attitudes by taking deliberative responsibility on the basis of *elusive* reasons, i.e., considerations that, unbeknownst to her, are biased and manipulated [4].

The problem is not that *S* self-ascribes a belief about *p* that she does not *presently* have, or that her belief about *p* is false. The trouble is rather that the self-ascription, although authored in the sense required by *AA*, is not a reliable guide to *S*'s future conduct. This is a problem, according to Lawlor, because she takes this reliability to be part of the idea of first-person authority. As Lawlor (2003, Section 9) says, "One's self-ascriptions are authoritative, in part anyway, because they are reliable expressions of those attitudes that govern further choices and behavior." The self-ascriptions based on elusive reasons lack what Lawlor (2003, Section 9) calls *agential authority*, since they are "no guide to how the person will ultimately choose, act, or feel about her choices and actions in the future."

The results of the psychological experiments presented by Lawlor are plausible and I have no reason to dispute them. What I want to challenge, instead, is Lawlor's claim that these and similar cases show that there is something amiss with *AA*. At the heart of her criticism is the claim that *AA*'s focus on authorship obscures the importance of agential authority in the account of the special, first-personal features of self-ascriptions. According to Lawlor, *AA* overlooks the significance of the relation between words, attitudes, and deeds that is the basis of agential authority

(and, as such, central to the proper account of first-person authority). Contrary to this view, I will argue that *AA* is sensitive to the requirements of agential authority, but it does not take agential authority to be the key to first-authority, which is instead primarily a matter of authorship.

3. Agential authority

There are two different aspects of agential authority, which should be discussed separately. First, the question of the correspondence between the subject's self-ascriptions and her actions; second, the issue of the temporal stability of the subject's attitudes and their effects her future conduct.

It is part of the nature of attitudes such as beliefs and intentions that they are connected with the subject's conduct. The exact nature of this connection is controversial, but for present purposes we can ignore the disputes concerning the details of this connection. The important point here is only that, because of the nature of these attitudes, a subject *S* cannot be taken to have and self-ascribe these attitudes unless she is going to behave in ways that are consonant with the attitudes in questions.

Any account of first-person authority that does not respect the connection between the subject's self-ascriptions and her conduct is thus to be rejected. But this criticism cannot be moved against *AA*. *AA* does not suggest that the subject can author attitudes while ignoring the nature of these attitudes, including their relations to the subject's conduct. *S* cannot take deliberative responsibility for the belief that *p* if she does not understand that this attitude is expected to affect her future conduct in specific ways. Likewise, *S* cannot self-ascribe this belief if her dispositions are such that the belief is not expected to affect her conduct in the ways appropriate to her circumstances. If we suspect that *S* is self-ascribing an attitude with no understanding or inclination to connect it with the appropriate conduct, we are justified in calling the self-ascription hollow and empty. In cases of this sort, the subject is confused about what she is self-ascribing, if she is self-ascribing anything at all. These self-ascriptions would indeed carry no agential authority.

This is, however, something fully acknowledged by *AA*. The authorship at stake in *AA* is not that of the gratuitous creation of attitudes that either disregards or defies the metaphysical, conceptual, and logical constraints imposed by the nature of the authored attitudes, including the constraints on the subject's future conduct [5]. If the subject violates any of these constraints, she can only *pretend* to be taking deliberative responsibility for the attitude in question. Her alleged self-ascription would sound hollow and empty because it would be a *spurious* self-ascription. This is exactly what *AA* says about these cases. The emptiness of spurious self-ascriptions is not a challenge to *AA*.

This does not seem to be, however, the trouble with agential authority specifically raised by the experiments presented by Lawlor. In these experiments, the self-ascriptions are taken to be genuine. Nonetheless, Lawlor says that they sound hollow. What they lack must thus be something related to the second component of

agential authority, i.e., the *temporal stability* of the self-ascribed attitude and how this stability is connected with future conduct.

4. Temporal stability

First-person authority is characteristic of self-ascriptions of *present* attitudes. Nonetheless, the self-ascribed attitudes are usually expected to influence or determine the subject's future conduct in specific ways. *Under normal circumstances*, *S*'s sincere self-ascriptions are reliable guides to *S*'s future behavior (conditional on the occurrence of the appropriate conditions). To this extent, sincere and genuine self-ascriptions carry agential authority. No plausible account of first-person authority can utterly deny or ignore this connection between self-ascriptions and conduct. As shown above, however, *AA* cannot be faulted on this count. *AA* might still be liable, nonetheless, to a weaker form of criticism, that is, that *some* genuine and sincere self-ascriptions are not reliable guides to future conduct (and thus lack agential authority) *even if* they are perfectly in order as far as the authority granted by *AA* (that is, even if the subject is taking full deliberative responsibility for the self-ascribed attitudes). The problem with *AA* would thus be that it makes attributions of first-person authority too liberal.

Is this kind of agential authority necessary to the idea of first-person authority? One might agree that central instances of authoritative self-ascriptions are reliable guides to future conduct, but nonetheless deny that this is necessary to first-person authority. This seems true especially if one considers what happens to the connection between self-ascribed attitudes and future conduct under a certain set of unusual, although by no means rare, circumstances. More precisely, there are three kinds of circumstances that prevent a presently self-ascribed attitude from determining the subject's future conduct in the way appropriate for that attitude. When this is known to happen, the self-ascription cannot be taken to be a reliable guide to future conduct. The self-ascription lacks agential authority. Nonetheless, as I am about to argue, it still has first-person authority.

Let's consider these three kinds of circumstances with reference to the case of *S*'s self-ascription of the belief that *p* and its connection to *S*'s action α at the future time *f*, where α is the action that is rationally expected of *S* at *f* given her believing that *p*.

- (1) At or before *f*, conditions unexpectedly change such that either doing α at *f* is no longer rational, or *S* at *f* no longer has the capacity or the opportunity to do α .
- (2) At *f*, α is still the rational thing to do and *S* has the opportunity to do it, but *S*'s rational capacities fail and she is no longer able to see that α is required of her: either *S* at *f* no longer sees the rational connection between *p* and α (say, because she is akratic or self-deceived), or she has unjustifiably changed her mind about *p* and she no longer believes it.
- (3) Neither (1) or (2) is the case, but the fully rational subject *S* realizes at *f* that her earlier belief must be rejected because it was based on either faulty

reasoning or inaccurate information. Therefore, S no longer has reason to do α .

What are the effects of these circumstances on the authority of present self-ascriptions and the tenability of AA ? Consider first cases of kinds (1) and (2), where the lack of agential authority is due either to changes in external circumstances or to the failure of the subject's rationality. The prospect that S is not going to do α under these circumstances appears to be immaterial to the special, first-person authority of S 's present self-ascription. Whether S indeed believes that p depends on her acknowledging the reasons in support of p , not on whether future conditions will make both possible and rational for her to act on the basis of this belief. S is concerned with the justification of her present attitude, not with either prediction or explanation [6]. Moreover, any prediction about S 's conduct at f must be based on a *prior* and *independent* determination of S 's belief about p . To this extent, the authority of authorship has priority over agential authority: S must self-ascribe an attitude with the authority of authorship before anyone (S included) can use that ascription as a guide to S 's future conduct.

The fact that either (1) or (2) might occur in the future does not affect the authority of S 's *present* belief that p , at least when S does not expect the future countervailing circumstances. These circumstances, however, affect the authority of a correlated self-ascription. On the basis of her belief that p , S might come to believe that she will do α at f . Because of the countervailing circumstances, this prediction is going to turn out false. Thus S 's belief that she will do α lacks some kind of authority. For it is false and it shows S not to be knowledgeable about her future α -ing. This is so even if both the belief that p and the self-ascriptions of this belief are true. Under these circumstances, neither of S 's self-ascriptions is a reliable guide as to S 's future α -ing, and thus should not be used as a guide by any observer who knows better than S about the future occurrence of the countervailing circumstances. This, however, does not make S 's self-ascriptions hollow or empty as self-ascriptions of attitudes. The *distinctive* first-person authority of these self-ascriptions concerns whether the subject *has* the attitudes, i.e. whether she takes responsibility for them, not whether the self-ascribed attitude is both correct and a reliable guide to future conduct.

What if S knows about the future countervailing circumstances? In this case, S is no longer justified in expecting her future α -ing. This does not affect, however, the authority of her self-ascription of the belief that p . What would be in trouble is rather the authority of the belief that she is going to do α . Given the knowledge of the countervailing circumstances, the subject cannot insist on claiming that she will do α . Were she to do so, this would suggest that she simply does not understand what she is self-ascribing. In this case, there is indeed something hollow or empty with her self-ascription, but this is not because the self-ascription is not a guide to her future behavior. Rather, it is because the self-ascription is spurious. This is not a problem for AA , however, since—as shown above— AA does not grant any authority to spurious self-ascriptions.

The hollowness of spurious self-ascriptions should not be confused with the

hollowness of the self-ascriptions made by a subject who is prone to akrasia or self-deception. This subject cannot be expected to act on her self-ascribed attitude. Her self-ascription lacks agential authority because it cannot be used to predict the subject's future conduct. Nonetheless, this does not undermine *AA*, since the self-ascription is correct *as a self-ascription*. The subject *does* believe that *p*, and she believes so because she authors the belief by taking deliberative responsibility for it. The trouble in this case is that she is not acting or going to act on the reasons that, as far as this self-ascription is concerned, she *endorses*. She is not simply going through the motions of deliberation; on the contrary, she is sincerely taking responsibility for the attitude. The problem is rather to be found in the connection between this sincere endorsement and the subject's other endorsements and actions, hence the akrasia or self-deception. If we were to deny authority to the self-ascription, we would not be able to understand the nature of the problem. Akrasia and self-deception are troublesome exactly because the subject fails to act on the attitudes that she has sincerely endorsed. It is only because her self-ascriptions are authoritative—in the sense of authorship—that something should be done to remove the causes of akrasia and self-deception, and thereby make the original self-ascription also agentially authoritative.

The lack of agential authority in the cases presented so far is not a valid challenge to *AA*. This is because *AA* has no difficulty acknowledging the existence of self-ascriptions such that either (a) the self-ascribed attitudes are incorrect or false; or (b) the self-ascribed attitudes are not reliable guides to future conduct; or (c) the self-ascriptions are apparent, confused, or spurious. These are common occurrences and no plausible theory of first-person authority would deny them. The lesson to be drawn by looking at these cases is not to reject the authorship account but to exert caution in taking self-ascriptions as guides to future conduct. But this does not make the self-ascriptions any less authoritative as ascriptions of attitudes to the subject.

This is not to say that the common sense view of first-person authority is mistaken in demanding that self-ascriptions be reliable guides to future conduct. *AA* does not deny that *under normal conditions* self-ascriptions are such a guide. Nor does it deny the conceptual connection between *p* and α -ing such that *S* is expected to do α if all of the following is true: (i) *p* is the case, (ii) *S* believes that *p*, (iii) *S* self-ascribes the belief that *p*, and (iv) there are no countervailing circumstances to *S*'s α -ing. But there is no guarantee that *S*'s genuine and sincere self-ascription of the belief that *p* will induce *S* to do α . This is because the occurrence of the countervailing circumstances is independent on whether *S* takes responsibility for *p*, which is all that matters as far as the first-person authority of the self-ascription is concerned.

5. Elusive reasons

A similar conclusion can be reached when considering the scenario (3). If *S* takes responsibility for a certain attitude under conditions that make the attitude either unjustified or incorrect, one should be wary of using this self-ascription as a guide to *S*'s future conduct. This is because *S* is a *rational* subject. As such, she might

come to realize that her original deliberation was flawed, and thereby change her attitude according to her better judgment. The psychological experiments presented by Lawlor seem to fit a scenario of this kind, since the subjects in the experiments are induced to sincerely self-ascribe attitudes on the basis of considerations that, unbeknownst to them, are manipulated and biased. It is thus to be expected that, if the subjects are ever to realize the existence of bias and manipulation, they will give up the attitude that they originally self-ascribed.

However, from the mere fact that an attitude is unjustified and incorrect (including its being based on manipulated and biased information), it does not follow that the self-ascription of the attitude cannot be a good guide to the subject's future conduct. We are familiar with cases of self-ascribed beliefs that, despite being unjustified and incorrect, are both stable over time and very reliable guides to the subject's future conduct. Do these self-ascriptions sound empty and hollow? Not necessarily. If because of elusive reasons *S* has the incorrect belief that *p*, then *S* does not believe the content *q* that she would have believed if she had considered the question under ideal conditions (which include the lack of biased and manipulated information). But this ideal belief is not what *S* *really* believes. The first-person authority concerns the real, not the ideal belief. An external observer might know better than *S* what *S* should believe if she were fully rational and under ideal deliberative conditions. Nevertheless, such observer must take *S*'s self-ascriptions as the authoritative guide about *S*'s conduct. For as long as *S* is unaware of the incorrectness of her belief, it is the actual self-ascription, not the ideal one, that makes a difference to *S*'s conduct. In cases of this sort, the self-ascription is authoritative both in the authorship and the agential senses, even if the belief is based on elusive reasons. These cases, therefore, present no challenge to *AA*.

The only cases of type (3) in which agential authority is missing are those in which it is to be expected that a *minimal* exercise of reflective and critical capacities on *S*'s part is sufficient to make *S* realize that the belief that she originally self-ascribed was unjustified and incorrect. If so, *S*'s is going to reject the original belief and she is going to self-ascribe a different one. In this case, there is a sense in which *S*'s original self-ascription is not indicative of her "real" belief and thus hollow and empty. What this means is that we know that *S*'s present self-ascription is very soon to be rejected. As soon as *S* reflects on her belief and the justification for it, she is going to recognize that the belief is unjustified and thereby reject it. Hence, the self-ascription is not a good guide to her conduct because, by the time she has to act on the belief, she will no longer hold it. This does not mean, however, that the original self-ascription is incorrect. This self-ascription *correctly* indicates what *S* believes *at the time* of the self-ascription. The subject is still in a privileged position with respect to her *present* attitudes. This is so for the reasons indicated by *AA*: what the subject believes is what the subject is taking responsibility for; it is not what she is going to take responsibility for once she becomes aware of the elusive nature of reasons that supported the conclusion of her original deliberation.

Before considering how these claims affect the plausibility of *AA* let me bring attention to the fact that the challenge based on elusive reasons does not really require appeal to the findings of social psychology. As long as a belief is based on

reasons that are manipulated and biased, and such that only a minimal level of reflection would bring this out, then the self-ascription of that belief is hollow in Lawlor's sense. We are familiar with many ordinary situations of this kind. The appeal to the psychological experiments mentioned by Lawlor is unnecessary to present her challenge to *AA*. The experiments show a particular way in which elusive reasons can be generated, but the challenge does not depend on the specific way in which the subject comes to take the elusive reasons seriously. The hollowness of the self-ascriptions is due to the manipulated and biased nature of the considerations on which the attitude is justified, not on the specific ways in which the bias and manipulation are generated.

Do the (3) cases challenge *AA*? In the discussion of the (1) and (2) cases, I argued that *AA* acknowledges the importance of agential authority, but it is not undermined by the lack of such authority. The same is true of the (3) cases. Actually, the (3) cases offer indirect support for *AA*, since—as I am about to show—*AA* offers a most convincing explanation of why agential authority is lacking in these cases.

The lack of agential authority in (3) depends on the temporal instability of the self-ascription. The source of this instability is the very phenomenon that is at the core of *AA*, namely, that (i) the subject takes responsibility for the self-ascribed attitude; and (ii) she does so by endorsing reasons that support the correctness of the attitude. The instability of a self-ascribed attitude based on elusive reasons can be expected because the responsibility that the subject takes in acquiring the self-ascribed attitude is not discharged only at the time of the original self-ascription. The self-ascription is expressive of *S*'s recognition of the reasons in support of the self-ascribed attitude; but *S* knows that she is fallible. *S* knows that she might be mistaken in her original endorsement. This means that she is ready to change both the attitude and its self-ascription if in the future she discovers that she made a mistake in the original deliberation. Deliberative responsibility is not just a matter of the original authoring of the attitude, but also of a *readiness* to change the attitude in light of any subsequent change in the recognition of the reasons supporting it. As soon as *S* finds out that her original reasoning in support of the belief that *p* was mistaken or invalid, she is expected *ipso facto* to drop this belief and its self-ascription [7].

It is because of this readiness that one anticipates the instability of any attitude (and its self-ascriptions) based on elusive reasons. Soon after the original incorrect endorsement, *S* is expected to recognize that her original endorsement was incorrect and to revise her attitudes and self-ascriptions accordingly. This means that the original self-ascription lacks agential authority, since it cannot be expected to determine *S*'s future conduct. Nevertheless, the lack of agential authority can be explained in terms of the authority of authorship. It is a product of such authority: it is only because *S* authors the attitude by taking deliberative responsibility for it that one can anticipate the attitude's instability on the face of the elusive nature of the original considerations [8]. Hence, the (3) cases present no challenge to *AA* [9].

To sum up, *AA* is particularly well suited to accounting for the distinctive features of agential authority, including its failures. This is because, according to

AA, the self-ascription of judgment-sensitive attitudes—e.g. beliefs and intentions—is based on the deliberative recognition of the reasons that justify these attitudes (see Moran, 2001, p. 115). Self-ascriptions are reliable guides to future rational behavior because they are the manifestation of the (fallible) recognition of the reasons supporting the self-ascribed attitudes. It is because these reasons are endorsed first-personally by a *rational* subject that one expects the subject to behave in conformity with these reasons in the future, *provided* that the conditions are appropriate and the subject's rational capacities are in order (that is, provided that neither (1), (2), nor (3) occurs).

Taking deliberative responsibility for an attitude, however, cannot guarantee that the countervailing conditions (1), (2), or (3) are not going to occur in the future. This is something that is not under *S*'s control in her acknowledging the reasons that support the self-ascribed attitude [10]. This is why there is no assurance that an authoritative self-ascription is also going to be a guide to future conduct, that the authority of authorship is accompanied by agential authority. *Ideally*, a self-ascription should both be true and a reliable indicator of future conduct, but achieving the former is no guarantee of the latter. At the core of first-person authority is only the authority of authorship. This authority, however, is exerted with the regulative ideal of securing the other kinds of authority, i.e. in view of the acquisition of attitudes that are both correct and reliable guides to future conduct.

6. Amending the authorship account

Because of the self-ascribed attitudes might be either incorrect or unreliable guides to future conduct, we should be careful in our self-ascriptions. We should make sure, for instance, that the self-ascribed attitudes are not based on elusive reasons [11]. Does this mean that I am endorsing the sort of emendation to *AA* that Lawlor presents as a possible, but ultimately unconvincing, response to the problems generated by elusive reasons? According to her (2003, Section 9), a defender of *AA* might try to avoid these problems by suggesting that self-ascriptions are authoritative *only if* the subject is a *diligent* deliberator, i.e. a deliberator who is “disposed to test one's reasons, and perhaps even develop some skill in recognizing biasing factors on one's judgments.”

This is not what I am suggesting. In my view, *AA* grants authority also to relatively negligent deliberators, provided that they are taking some sort of deliberative responsibility for their attitudes. I maintain, however, that some degree of diligence is normally expected of rational deliberators. Diligence is one of the virtues that must be exerted in order to approximate the model of ideal deliberation. Hence, the requirement of diligence in deliberation is already built into the very idea of being the author of an attitude by taking the deliberative stance. For one cannot be said to be truly deliberating if one were not somehow sensitive to the possible deviations from the normative standards of deliberation. However, some deliberators are more reflective, attentive, careful, and diligent than others. This means that the outcomes of their deliberations are closer to securing the ideal condition; i.e. the condition in which the agent can speak with authority about her attitudes in both the

authorship and the agential sense, and being authoritative in her attitudes because the attitudes are correct and justified. But *AA* does not suggest that there is genuine first-person authority only in the ideal situation. Authorship is the foundation of first-person authority, and this authority is present even if, in exerting it, *S* falls short of the ideal of authoring only correct and stable attitudes.

My appeal to diligence should not to be interpreted as the emendation to *AA* discussed by Lawlor. For it is not really an *emendation* to *AA*. This move, however, does not seem sufficient to avoid Lawlor's fundamental criticism, since it fails to address what Lawlor (2003, Section 9) calls the central problem.

The central problem is not that subjects occasionally err, or are less than diligent in their deliberation about what attitudes to have. The central problem is that, however well they deliberate, subjects can deliberate themselves into attitudes that just don't govern their behavior. It is this aspect of first-person authority that authorship accounts miss, even with the proposed modifications.

This quote reveals that Lawlor is not just looking for an account of first-person authority that does justice to the connection between self-ascriptions, attitudes, and behaviors. She insists on a stronger requirement, i.e. that the account *guarantees* that the self-ascribed attitudes govern behavior [12]. Agential authority is for her the real key to first-person authority.

We are committed to the idea that first-person authority *derives* from reliable correlations between what one is inclined to say and what one is inclined to do, between the attitudes one sincerely asserts and the attitudes one acts upon (Lawlor, 2003, Section 9; emphasis mine).

As I have argued in this paper, this is too strong a requirement to impose on accounts of first-person authority. Although I agree with Lawlor that securing agential authority is important, this is only a regulative ideal of deliberation, not a necessary condition of it. This is all that is required in order to do justice to the connections between words, attitudes, and deeds. To this extent, nothing can be objected to *AA*. Lawlor is right however in claiming that, *if* we want the connections to take the form of *guaranteed* reliable correlations, we must look for something other than *AA*. But I see no reason why we should impose this more stringent requirement on an account of first-person authority. If we were to do so, we would lose sight of the really important issue, i.e. the distinctive first-personal features of self-ascriptions.

The key to *AA* is the priority it gives to the authority of authorship over the other kinds of authority. This is crucial, since the distinctive *first-personal* features of both the attitudes and their self-ascriptions depend on authorship alone. The other kinds of authority (those produced by correct, stable, and reliable attitudes) are important in the sense that a rational subject aspires to them in taking responsibility for her attitudes. But they are not the sources of distinctively first-personal authority. In order for the subject to speak with this authority, it is both necessary and sufficient that, as *AA* claims, she *authors* her attitudes and their self-ascriptions by the exercise of her capacity for rational deliberation.

Acknowledgements

The author wishes to thank Richard Moran and Carla Bagnoli for helpful conversations.

Notes

- [1] See Moran (2001, pp. 21–24) for a criticism of Crispin Wright’s social concession account.
- [2] On the authority of authorship, see Moran (2001, pp. 63, 92, 113).
- [3] A self-ascription of an attitude must therefore answer to the requirements of the correctness of the self-ascribed attitude. The subject is not at liberty of self-ascribing attitudes in a voluntaristic fashion. Against the confusion of the authorship account with voluntarism, see Moran (2001, pp. 64, 114–120, 127, 140).
- [4] For a detailed description of these works, see Lawlor (2003, Section 8).
- [5] See, for instance, Moran (2001, pp. 87–88), where he discusses how a self-ascription of an attitude is answerable to psychological evidence about the attitude even if the self-ascription is the product of a deliberation about the reasons supporting the attitude rather than an investigation into the psychological evidence.
- [6] On the distinction between justifying and explanatory reasons and its relation to first-person authority, see Moran (2001, pp. 128–130).
- [7] On the agent’s readiness to change her attitudes in response to the changes in her assessments, see Moran (2001, pp. 115, 117–118, 146).
- [8] The instability that is anticipated in cases of this sort is the product of the subject’s primary concern with the truth of her beliefs on the face of stable external circumstances. The instability can thus be anticipated only by taking the subject’s perspective. More precisely, only by taking the perspective of a properly idealized subject, given that the actual subject is not able to anticipate her change of mind since she is still under the mistaken impression of the validity of the elusive reasons. For a comparison between the subject’s and the interpreter’s perspectives in the prediction and explanation of behavior, see Moran (2001, pp. 129–130).
- [9] What about, however, the external observer who knows of the elusive reasons in advance of *S* and thus can be said to know better than *S* what her real attitude is going to be? Doesn’t this undermine *AA*, given that *S*’s attitude can be determined in *advance* of her authoring it? Not so. The observer is said to know *S*’s real attitude only because he is able to *anticipate* what *S* is going to endorse in the first-person when she is under better deliberative conditions. The privileged position of the observer is still parasitic on the subject’s future exercise of the authority of authorship, which is the ultimate source of the attitude and its self-ascription. If it were not for the subject’s authorship, there would be no attitude of whose existence and content the observer could be said to know better (see Moran, 2001, p. 129).
- [10] This is not to say that the subject might exert control over the future occurrence of any of the countervailing conditions. The subject might, for instance, be partially in control of her future akrasia or self-deception if some kind of psychological therapy were available that could cure her of either. My point here is only that whatever control the subject might have over the future occurrences of the countervailing conditions, this is not the control that she is exerting *in* taking deliberative responsibility for the future belief about *p*. This is so, at least, for the case of a content *p* which is not about *S*’s future behavior. Things are more complicated if the belief is about *S*’s future behavior, since she cannot be said to be taking responsibility for the belief that she is going to do β at *f* while she knows that there are countervailing conditions to her future β -ing. This consideration is especially important in the case of the self-ascription of intentions about future actions. The account I offer in the main text would thus need to be slightly modified in order to apply to the self-ascription of intentions. But the conclusion would not change.
- [11] Likewise, we should rule out conditions (1) and (2), if we want to make a reliable prediction of our future conduct on the basis of our present self-ascriptions.

- [12] This is clear, for instance, in what she says to justify the rejection of the emendation to *AA*, namely, that the emendation gives “no guarantee that care in deliberation will result in attitudes that govern how a person will ultimately choose, act or feel. Even in cases of the most diligent deliberation, one’s readiness to take up deliberative questions about what one’s attitudes are to be is not enough to rescue one’s authority” (Lawlor, 2003, Section 9).

References

- LAWLOR, K. (2003). Elusive reasons: a problem for first-person authority. *Philosophical Psychology*, 16, 549–564.
- MORAN, R. (2001). *Authority and estrangement: an essay on self-knowledge*. Princeton, NJ: Princeton University Press.
- SELIGMAN, C., FAZIO, R.H. & ZANNA, M.P. (1980). Effects of salience of extrinsic rewards on liking and loving. *Journal of Personality and Social Psychology*, 38, 453–460.
- WILSON, T.D., HODGES, S.D. & LAFLEUR, S.J. (1995). Effects of introspecting about reasons: inferring attitudes from accessible thoughts. *Journal of Personality and Social Psychology*, 69, 16–28.

