THE EPISTEMIC BENEFITS OF REASON GIVING

Forthcoming in *Theory & Psychology*

Lisa Bortolotti

Abstract

There is an apparent tension in current accounts of the relationship between reason giving and self knowledge. Philosophers like Richard Moran (2001) claim that deliberation and justification can give rise to first-person authority over the attitudes that subjects form or defend on the basis of what they take to be their best reasons. On the other hand, the psychological evidence on the introspection effects and the literature on elusive reasons suggest that engaging in explicit deliberation or justification leads subjects to report attitudes that are not consistent with their previous attitudes or with their future behavior. On the basis of these findings, Tim Wilson (2002) argues that analyzing reasons compromises self knowledge. I shall defend a realistic account of the effects of reason giving which is compatible with the empirical findings on introspection and also with the claim that deliberation and justification have epistemic benefits.

Keywords

Rationality, authority, authorship, self knowledge, introspection, reflection, deliberation, justification, belief.

1. Giving reasons and gaining authority

Does reason giving contribute to or detract from self knowledge? In this first section I shall introduce the notion of authorship, and identify a connection between endorsing the content of one's attitudes with reasons and having first-person authority over those attitudes. In section two I shall review objections to the view that reason giving contributes to self knowledge. These objections are based on the psychological evidence on the introspection effect and elusive reasons. In section three I shall argue that the practice of reason giving has epistemic benefits, although there is no good reason to believe that being the author of one's attitudes is sufficient for the rationality[1] and stability of the reported attitudes or for knowledge of the psychological mechanisms causally responsible for the formation of those attitudes.

1.1. Self knowledge and reason giving

In both philosophy and commonsense we distinguish between two ways of characterizing the asymmetry between first and third person. The first person has a privileged claim to the knowledge of her own attitudes because she typically has direct access to her attitudes, and because she exercises some control over what she comes to believe, feel, choose, and so on.[2] The standard ways in which a subject gains control over her attitudes are deliberation (when an attitude is formed on the basis of reasons that are regarded by the subject as her best reasons for endorsing its content), and justification (when a subject defends the content of her attitude on the basis of reasons that are regarded by the subject as her best reasons).

In the philosophical literature, it has been argued that the processes that make it possible for a subject to gain control over her attitudes give that subject a claim to self knowledge (e.g. Moran 2001). In the psychological literature, the consensus is that people have no genuine control over the attitudes they report. Even when they seemingly engage in deliberation and justification, and analyze reasons in favor of their attitudes, they do so in ways that compromise rather than support their claims to self knowledge (e.g. Wilson 2002). In this paper I shall attempt to explain the apparent tension between these two approaches to the relationship between reason giving and self knowledge, and come to a conciliatory position, according to which reason giving does not produce self knowledge, but has significant epistemic benefits.

In standard characterizations of self knowledge, it is assumed that a subject can ascribe mental content to herself in at least two ways, as a result of direct epistemic access to the content of her mental states or as a result of inferences based on evidence about her own behavior. When she has direct epistemic access to her mental states, she does not need to rely on behavioral evidence to work out what the content of those states is and therefore she has a *prima facie* advantage over external observers (e.g. Burge 1988; Heil 1988; Peacocke 1998). Even when the subject lacks direct epistemic access, she can often rely on more and better evidence about her own behavior than a third person, and generally she can make more successful inferences from her behavior to the content of her mental states (e.g. Ryle 1949). I shall call 'epistemic' the route to first-person authority that is based on the subject's capacity to obtain better epistemic access to her mental states than a third person, either by introspection or by inference.

The other way of coming to know one's own attitudes can be illustrated by introducing and distinguishing the notions of ownership and authorship of thoughts. One is usually in a condition to self ascribe those thoughts that can be accessed directly or inferred from one's own behavior. I call *ownership* one's capacity to acknowledge a thought as one's own which leads to ascribing the thought to oneself. Ownership is achieved on the basis of (direct or mediated) epistemic access to the content of one's own thoughts. But there is another route to self ascription which involves other capacities. In some circumstances, one can ascribe a thought to oneself on the basis of an act of *authorship*. One is the author of a thought, say, if one forms it or justifies it on the basis of what one takes to be one's best reasons.

Consider the belief that I need to cross two busy roundabouts on my way to work. There is very little about that belief that is *up to me*. If challenged, I can support the belief by appealing to the accuracy of my perceptual experiences and the reliability of my memory. On an everyday basis, I can act in a way that is compatible with my having that belief. But in no other significant way can I be said to endorse the belief that there are two roundabouts on my way to work. I manifest my authorship potential to a greater extent with respect to those beliefs whose content I am prepared to defend on the basis of reasons, such as opinions, or with respect to other attitudes, such as intentions and motivated desires (see Carman 2003). The belief that the debt of developing countries should be cancelled, for instance, is the type of belief that I can endorse with reasons. I can argue for the claim that cancelling the debt would have significant benefits for people in developing countries by allowing their governments to invest more resources in education. The belief is authored by me, because it is up to me what to believe, and I take

responsibility over its content by having formed it or defended it on the basis of what I consider to be my best reasons.

Moran argues that authorship provides a route to first-person authority that is distinct from the epistemic one. In order to appreciate the distinction we can reflect upon the common practice of ascribing beliefs on the basis of observed behavior. There are circumstances in which it makes sense to say that, as interpreters of the behavior of others, we might come to know what their beliefs are better than they do themselves. Even if one has direct epistemic access to one's beliefs or richer evidence about one's own behavior, one is vulnerable to self deception or self attribution biases, or may have beliefs that are not always open to introspection. But not even a very competent interpreter can ever author someone else's belief, because authorship requires the capacity to endorse the belief on the basis of what one takes to be one's best reasons. If authorship of a belief gives rise to first-person authority over that belief (which is the object of this paper), this authority does not derive from a privileged form of epistemic access, but rather from a series of *actions* that one can perform with respect to the content of that belief. The subject can deliberate, justify, endorse, and manifest a commitment.[3] I shall call 'agential' the route to first-person authority which is manifested via authorship.

Moran argues that authorship does give rise to first-person authority and that agential authority is an important dimension of self knowledge.

> If it is possible for a person to answer a deliberative question about his belief at all, this involves assuming an authority over, and a responsibility for, what his belief actually is. Thus a person able to exercise this capacity is in a position to declare what his belief is by reflection on the reasons in favor of that belief,

> rather than by examination of the psychological evidence. In this way […] avowal can be seen as an expression of genuine self-knowledge. (Moran 2004a, 425)

1.2. Agential and epistemic routes to first-person authority

One reason to explore the distinction between epistemic and agential authority is that they can come apart – that is, one can have epistemic authority without having agential authority. As I have already suggested with some examples in the previous section, some attitudes are better suited to being endorsed than others. So agential authority will not apply (or will apply to a lesser degree) to mental states whose content cannot be endorsed on the basis of reasons. But let's concentrate on those attitudes that are usually endorsed on the understanding that reasons could be provided in favor of their content. In these cases, when the epistemic route to authority is intact, but the agential route is compromised, it would seem that epistemic access is not sufficient for self knowledge.

It is possible to be aware of the content of a thought in a first-personal way, thanks to direct epistemic access, and yet fail to have a sense of ownership and authorship with respect to that thought. A dramatic example of alienation from one's thoughts consists in the experience of some people who suffer from thought insertion. The subject reports the delusion that others have inserted thoughts in her mind. She is unable to self ascribe the thoughts she can access first-personally and often describes her experience as 'others thinking in her head', as if her head was an instrument others could use to form or express thoughts (Sims 2003, ch. 9). It would seem that access to the content of the thoughts is intact but both the sense of ownership and authorship are lost with respect to

the 'inserted' thoughts, because the subject does not acknowledge the thought as her own and typically the subject does not endorse the content of the thought on the basis of reasons (Bortolotti and Broome forthcoming).[4] In cases such as this, when a subject maintains direct epistemic access over a thought but cannot ascribe the thought to herself, or provide reasons to endorse its content, we question the extent to which the subject has self knowledge. We usually describe the phenomenon of thought insertion as a paradigmatic failure of self knowledge: not only does the subject experience a violation of her personal boundaries, but she also fails to recognize that the thought she can access first-personally is *her own* thought.

Reflection on phenomena such as thought insertion leads us to suppose that, in some circumstances, acknowledgment of a thought as one's own and endorsement via reason giving make a significant contribution to self knowledge. The claim could be put more strongly: it is plausible to argue that, for some thoughts, when a thought is neither ascribed to oneself nor endorsed via reason giving, the subject does not have knowledge of having it. One can give reasons for having the belief that we should cancel the debt of developing countries; the desire to move to Spain; the preference for musicals over operas; the decision to withdraw from a competition. Where ownership and authorship of these attitudes are missing, the subjects' failure to manifest agential authority undermines attributions of self knowledge to them.

The claim that reason giving provides an additional route to first-person authority and is, in some circumstances, necessary for self knowledge is controversial. The relation between the capacity for reason giving and first-person authority has been attacked on the basis of the psychological evidence highlighting the limitations of introspection in

deliberation and justification. The interpretation of the evidence is that the reflective processes of analyzing reasons for attitudes undermine rather than promote self knowledge:

> It is common for people to analyze why they feel the way they do [...]. It is usually assumed that such reflection is beneficial, leading to greater insight about how one feels. We will argue the reverse; that is, that this type of self-analysis can mislead people about their attitudes, thereby lowering attitude-behavior correlations. (Wilson et. al. 1984, 5).

Here is a preliminary list of the phenomena observed:

(1) *Instability of attitudes* (*attitude/attitude inconsistency)*. When the reasons for the attitude are not available to introspection, analyzing reasons for the attitude causes one to 'change one's mind'.

(2) *Irrationality of attitudes.* The attitude reported after the reason giving exercise is not consistent with the attitudes previously reported, and is less optimal (e.g. worse choices with respect to expert judgment, increased vulnerability to evidence manipulation or inconsistencies).

(3) *Poor self prediction* (*attitude/behavior inconsistency)*. The attitude reported as a consequence of analyzing of reasons is not found to be representative of the one's future behavior.

(4) *Ignoring why*. When analyzing reasons for a reported attitude, one is blind to the 'real reasons' for that attitude and reason giving exercise does not provide any genuine insight into the workings of one's own mind.

Findings revealing an inconsistency or instability of attitudes, or an inconsistency between attitudes and behavior are taken to be evidence for a failure of first-person authority: when asked to think about reasons, subjects report attitudes that are not truly representative of their own viewpoint and that are not likely to be manifested in their future behavior. Thus, the evidence speaks against the quality and reliability of first-person ascriptions made when analyzing reasons.

The evidence that shows a failure in identifying the 'real reasons' for reported attitudes and the evidence disputing the quality of reported attitudes after reason giving are a challenge to the view that reason giving has epistemic advantages, although the link between the experimental results and the self-knowledge literature is more tenuous. The idea that people often lack awareness of the 'real reasons' for their attitudes is at the core of the view that judgments about ethics are due to affective responses and not to reasoning (Haidt 2001). Recent studies suggest that, when subjects engage in reasoning, they search for arguments that will support their already made judgments in an attempt to disguise moral dumbfounding.[5] This means that people are not aware of what really causes their moral judgments to be the way they are.

Whether the attitudes that are reported after engaging in reason giving are rational or optimal is relevant in the context of attitude shifts.[6] It rules out the possibility that subjects endorse different attitudes after thinking about reasons, because the search for reasons has made them realize what attitude is better supported by the evidence available

to them. If the attitude reported after engaging in reason giving is different from the previously reported one, less rational or less consistent with future behavior, then the implication seems to be that searching for reasons leads subjects astray.

We are faced with an apparent tension. Anomalous cases such as thought insertion lead us to appreciate the importance of ownership and authorship for self knowledge. But the standard interpretation of the psychological evidence on analyzing reasons, or on reasons-based judgments and evaluations (Wilson 2002; Hixon and Swann 1993; Lawlor 2003), invites us to see deliberation and justification as obstacles to the attainment of self knowledge.

Moran seems right that reason giving provides an additional route to the knowledge of the content of one's own attitudes and to the knowledge of one's having those attitudes. However, the psychological literature does suggest that when one introspects and thinks about reasons one can fail both to meet standards of rationality for the formation and justification of one's attitudes[7] and to identify correctly the mechanisms which are causally responsible for one's own conscious attitudes. In Moran's original account, the possibility of a discrepancy between the reasons for forming an attitude and the reasons provided for its *post-hoc* justification is underestimated. Moreover, the possibility of attitudes just 'popping up', that is, being formed in ways that are not accessible to introspection, and yet being justified at a later stage, is not sufficiently emphasized. Authorship is presented as causally efficacious in attitude formation, and as an exercise in rationality.

The psychological evidence tells us that not all instances of reason giving are aimed at making one's mind up about something. The reasons people offer for their attitudes are

sometimes just an attempt to mask the absence of justification with whatever reasons seem best to fit unmotivated reported attitudes seems to be epistemically blameworthy and scarcely conducive to (self) knowledge (Wilson 2002; Wegner 2002). But this is not the only role for *post-hoc* reason giving. One might have no introspective access to how one's attitudes were formed, and engage in a process of rationalization which allows one to *determine* or *discover* one's best reasons in support of the attitudes one happens to have. In this latter scenario, *post-hoc* reason giving is epistemically praiseworthy as it allows one to gain control over attitudes acquired in ways not transparent to introspection.

Not only does Moran tend to overestimate the extent to which one deliberates, but he describes authorship as an exercise in rationality. This could lead us to expect that both the processes and the outputs of deliberation and justification meet standards of rationality. The psychological evidence acts as a reminder that, in the real world, deliberation and justification are affected by the same limitations that apply to all other reasoning processes, and are hostage to selective memory and attention, poor thinking, evidence manipulation and attribution biases. These limitations constrain the extent to which the endorsed attitudes satisfy independent standards of rationality, and the extent to which they cohere with other attitudes and behavior. But these failures are not necessarily failures of self knowledge. My diagnosis of the apparent tension between the philosophical and the psychological literature is that the puzzling results of the introspection studies are due to a frequent breakdown of rationality, and to the biases affecting introspective reports in general. It is not reason giving in particular that is responsible for failures of self knowledge. As the case of thought insertion suggests, the

capacity for self ascribing and endorsing an attitude on the basis of reasons is, with respect to some attitudes, necessary for self knowledge.

The message often drawn from the introspection literature, e.g. that deliberation and justification get in the way of self insight and that we'd better go with the flow rather than stop and think, is not supported by the evidence. *Bad* instances of deliberation and justification make for the endorsement of attitudes that do not meet independent normative standards of rationality and might lead one to form false beliefs, about the world in general or about oneself. But this is hardly news and does not apply to *good* instances of deliberation and justification. All we need to concede to the standard interpretation of the psychological evidence is that first-person authority over the attitudes we can endorse with reasons does not guarantee the rationality and stability of the reported attitudes, and does not guarantee knowledge of the mechanisms responsible for the formation of those attitudes.

2. Authorship

Following what Moran says about the conditions for authorship of a belief, I author my belief that *p* if I am in a position to endorse that belief on the basis of reasons that I take to be my best reasons for *p*. Authorship does not require that the belief itself be true or rationally formed. It does not even require me to have better reasons in favor of, rather than against, the content of the belief. After all, I can be misled about what my best reasons are. But I must see myself as endorsing the belief for my best reasons in order to be its author. If this condition is not met, according to Moran (2001), the process leading

to the endorsement of the belief cannot be seen as a process of genuine deliberation or justification.

2.1. The deliberative stance

When I engage in deliberation or justification I do not report my belief or explain my behavior from a neutral stance; to some extent, I determine what the content of that belief is going to be and I exercise control over the belief I report. My stance towards the belief I author is different from your stance towards it, when you, as an interpreter, ascribe to me that belief on the basis of the observation of my behavior. In the interpretive stance, beliefs are ascribed to oneself or others in order to explain or predict behavior. In the deliberative stance, beliefs are justified on the basis of how things are according to the subject's best reasons and are expected to determine the subject's future behavior. This distinction is useful in order to map some of the asymmetries between first- and third-person knowledge.

The explanation of behavior provided in interpretation is not compromised if the beliefs ascribed are false or unreasonable by the standards of the interpreter. You can ascribe to me the intention to keep drinking in the pub until closing time, in order to explain instances of my behavior, but that does not mean that from your point of view my intention is a reasonable one. Suppose that I don't want to leave the pub early because I don't want disappoint my friends who only respect heavy drinkers. Suppose that I am actually bored to death by their unintelligible conversation, that I am tired, and that I have an important job interview in the morning. You can no doubt explain my behavior (that I

want to stay till closing time because I don't want to disappoint my friends) without being at all persuaded that my reasons are overall good reasons.

From my point of view as a deliberator, having reasons for an attitude is not sufficient to author that attitude. I need to believe that continuing to drink in the pub until closing time is the best thing to do, and I need to be able to support my intention to stay in the pub and keep drinking with reasons that I take to be my best reasons. It is possible that I am the kind of person who considers obtaining the approval of her friends more important than being in top form at a job interview.

But now suppose I do believe that the best option for me would be to go home and get some sleep, and I decide to stay in the pub nonetheless. Maybe I am weak-willed. I really should go home, I know that, but I cannot make myself leave. If asked to justify my decision the next day, when I get up with a serious hangover and puffy eyes, I can put together some reasons why staying in the pub was a good idea. If challenged, I can say that it was the best pub in town, that I have rare opportunities to spend quality time with my friends, and that I needed to relax before such a big day. In the context of justification leading up to authorship, how good I take my reasons to be does matter to whether I succeed in authoring my attitudes. But in the context of a half-hearted rationalization of my attitudes, it seems that standards have dropped. I myself know (or part of me does, depending on the preferred analysis of the situation) that the reasons I come up with are not my best reasons. In this latter version of the example, I no longer qualify as a deliberator or a justifier. The rules of the game of deliberation and justification don't permit it.

Haidt (2001) suggests that the majority of our attempts at arguing for the truth of our reported attitudes are not a causally efficacious exercise in deliberation, but a rationalization. He compares the job of a judge to that of a lawyer. In deliberation we are like judges who have to weigh up the evidence in favor and against a case, and make up our minds by that process. When we engage in *post-hoc* rationalization we are like lawyers trying to build a case, and we are no longer interested in the truth of the matter.

Does a subject have a claim to first-person authority over her attitude on the basis of the reasons she can offer in its support, after the attitude has been already formed and become available to her? It does depend on the details of the case. My suspicion is that very few cases will fall neatly into the two categories described by Haidt (2001): searching for the truth or deceiving oneself and others. If what makes the reason giving *post-hoc* is that the reasons adduced to support the attitude do not coincide with the reasons why the attitude was formed in the first place (or the causes of its formation, if no explicit deliberation was involved), then the conclusion that no first-person authority can ensue from the reason giving exercise seems rushed. If what makes the reason giving *post-hoc* is that the reasons adduced to support the attitude are not regarded by the subject herself as her best reasons, then the conclusion that no first-person authority can ensue from the reason giving exercise is more plausible.

The comparison between the context of interpretation and the context of deliberation (or first-person justification) is useful no matter how we adjudicate controversial cases of confabulation, self deception and weakness of will. It shows how by deliberating and justifying one can take responsibility over the content of one's own thoughts and how the first-person perspective is not exhausted by privileged epistemic access. Nothing in the

way in which the context of deliberation and justification are described suggests that reason giving needs to occur in ideal conditions in order to play a role in the acquisition of knowledge of one's own states.

2.2. Against authorship

Krista Lawlor (2003) argues that giving reason in support of an attitude does not give rise to first-person authority over that attitude: all the conditions for authorship can be satisfied, and yet the reported attitude can fail to accurately represent what the subject thinks. In order to reach this conclusion, she appeals to experimental evidence obtained in the course of a series of studies on the effects of introspective deliberation and justification. Results indicate that research participants are very likely to shift their commitments when they are asked to give reasons for their previously held attitudes (Wilson and Hodges 1994); they are more vulnerable to the effects of evidence manipulation when they are asked to give reasons for their attitudes (Wilson et al. 1995); the quality of the decisions and predictions made on the basis of reflective introspection or reason giving is inferior to the quality of decisions and predictions made otherwise (Wilson and Schooler 1991; Halberstadt and Levine 1999); and when making a decision, looking for reasons slows down research participants' information processing and reduces their capacity to discriminate among important factors (Tordesillas 1999).

All these findings are interesting and somehow disconcerting, but, as Lawlor concedes, their impact on whether reason giving gives rise to first-person authority is not always clear. Research participants' initial attitudes often differ from their final attitudes after they have engaged in an analysis of reasons. This suggests that the analysis of the reasons

why one endorses a certain attitude is not always a faithful reconstruction of the reasons why that attitude was formed in the first place but it prompts an entirely new process of attitude formation (I will come back to this hypothesis). When the attitude endorsed after analyzing reasons is different from the attitude reported at the start, the discrepancy gives rise to a question about self insight: which of the two attitudes represents the participant's genuine standpoint?

The argument against authorship goes as follows.

- If giving reasons in support of one's attitudes contributes to first-person authority over those attitudes, then people who have the opportunity to analyze their reasons in support of their attitudes should enjoy first-person authority over those attitudes. [8]

- Research participants who are given the opportunity to analyze reasons for their attitudes systematically fail to exhibit first-person authority over those attitudes.

- Therefore, giving reasons in support of one's attitudes does NOT contribute to first-person authority over those attitudes.

The argument could also be presented as a comparative claim between the performance of research participants who have the opportunity to engage in introspective deliberation and justification, and research participants who don't.

- If giving reasons in support of one's attitudes contributes to first-person authority over those attitudes, then people who have the opportunity to

analyze their reasons in support of their attitudes should be more authoritative with respect to their attitude than people who are not given the same opportunity. [9]

- Research participants who are given the opportunity to analyze reasons for their attitudes are systematically less authoritative with respect to their attitudes than research participants who are not given the same opportunity.

- Therefore, giving reasons in support of one's attitudes does NOT contribute to first-person authority over those attitudes.

How is premise two justified in the arguments against authorship? When research participants who engaged in introspective deliberation or justification attempt to predict their own behavior on the basis of their reported attitudes, their predictions are less successful than those of a well-informed third person or of the self predictions of research participants who did not engage in introspective deliberation or justification. The interpretation of these findings is that reason giving undermines rather than promotes self knowledge.

To flesh out the arguments, I shall refer to a series of experiments that have been used as paradigmatic examples of failures of first-person authority in spite of reason giving. People dating for a few months are asked to give reasons why they are attracted to their partner, and then rate the commitment to the relationship and the likelihood that they will live together or get married in the future (Seligman et al. 1980). Participants' responses

are elicited via questioning. Guided by the formulation of the questions, some couples are invited to offer *intrinsic* reasons for their being together ('I date X because…'), whereas others are invited to offer *extrinsic*, more instrumental reasons ('I date X in order to…'). Research participants invited to give extrinsic reasons end up rating more negatively their attitudes towards their partners and tend not to predict living together or getting married in the future. They do not seem to realize that their reports are biased by the way in which the questioning was conducted: the fact that they give reasons for their attitudes does not guarantee that they report an attitude that accurately represents how they feel and what they think about the relationship.

Wilson and Kraft (1993) designed a similar study on attitudes towards one's partner (but without evidence manipulation). Participants were first asked to report their attitude towards their relationship, then they were divided in two groups, and finally asked for their evaluations again, and for a prediction about the future of the relationship. In one group, the intermediate task was to list reasons for the success or failure of the relationship. In the other group, they were given a different task. Results show that the participants who were asked for reasons for the state of their relationship experienced an attitude shift between the former and latter reports.[10] In a study where a follow-up interview was also included, Wilson et al. (1984) found that participants who evaluated their relationship and made a prediction without being asked for reasons made more successful predictions than the participants who were asked for reasons.

Lawlor (2003, 558) claims that in the studies on dating couples research participants lack first-person authority, because their attitudes are unstable and because participants are likely to behave towards their partners in a way that is not consistent with or explicable

by their reported attitudes. And yet, they were invited, in the course of the experiments, to think about the reasons for their reported attitudes. The suggestion is that the reason-giving exercise did not bring out the *real self* in the research participants: the attitudes endorsed as a consequence of analyzing reasons for dating were not manifested in the participants' subsequent behavior and did not reflect their commitment to the relationship.

In the original study there is something very unsettling about the pattern of the research participants' responses, and about the ease with which their responses were manipulated by the experimenters. However, it is not obvious that the participants experienced a failure of first-person authority over their attitudes. For Lawlor's argument to go through, we need to agree that research participants meet the conditions for authoring their reported attitudes and yet fail to gain authority over them. Lawlor is right that participants qualify for authorship. The evidence shows that they are not alienated from the attitudes they report and, when asked to do so, they give reasons for endorsing those attitudes which (presumably) they take to be their best reasons. We must conclude, then, that they meet the criteria for authorship with respect to the attitudes they report. Their reported attitudes are up to them and are 'answerable to their explicit thinking about the matter' (Moran 2001, 123). The next question is whether research participants fail to gain authority over their reported attitudes.

3. Conditions for first-person authority

The view that in the studies on dating couples research participants fail to exercise authority over their reported attitudes seems to assume that the following conditions need

to be satisfied for first-person authority. In order for a subject to have first-person authority over an attitude endorsed with reasons:

a) the attitude needs to be stable and consistent with other relevant attitudes and with future behavior;
b) the reasons for endorsing the attitude need to be also the reasons why the attitude was formed;
c) the subject needs to know why she has that attitude.

Let's examine these conditions.

3.1. Rationality and stability

In the original dating couples study, research participants form their attitudes and make their predictions on the basis of evidence for and against the likelihood of success of their relationship, but the evidence most accessible to them is constituted by their answers to the experimenter's questions. These answers were framed so as to promote a certain perspective on the relationship, either as something worthy in itself, or as something which is instrumental in order to achieve other goods. This manipulation in the experimental design leads research participants to a one-sided evaluation of their relationship and to a prediction that does not take into account all the factors which affect the success of the relationship.

The reasoning process by which they arrive at their attitude is far from rational: they can be charged with the violation of the principle of *total evidence*, which both Carnap and Davidson considered as a fundamental principle of rationality. According to this principle, all the relevant available information should be taken into account when

forming an attitude. This breakdown of rationality is largely responsible for the prediction that the participants' reported attitudes will not be consistent with their future behavior. The hypothesis here is that the participants' future behavior will be affected by all the important factors relevant to the success of their relationships, and not just by those factors which research participants have been asked to consider during their interview with the experimenter.

Notice that describing the processes leading to authorship as non-rational does not make the talk of authorship incoherent, because deliberation can give rise to authorship even when the reasons subjects have for deliberating are not good reasons, as long as they are what subjects take to be their best reasons. Whether authorship can result from non-rational deliberation or justification, though, is not sufficient to arrive at a verdict on whether the subjects lack self knowledge. We need to know whether *non-rational* deliberation or justification can be a source of first-person authority. The first step is to acknowledge that the capacity to report attitudes that are truly representative of one's state of mind and that will guide one's future behavior is limited. The second step is to resist the claim that deliberation and justification have to bear the sole or primary responsibility of these limitations. Deliberation and justification are reasoning processes, and they are affected by biases as any other reasoning process (Nisbett and Wilson 1977; Kahneman et al. 1982; Stanovich 1999).

Trivially, whether the processes of deliberation or justification are rational matters to whether the outputs of such processes are. The belief that a relationship is doomed to failure, due to the fact that no intrinsic reason for dating can be recalled at the time, is not a rationally formed belief. But this does not mean that the belief should not be ascribed to

oneself or genuinely endorsed as one's own. Even if the belief was not formed rationally, or no rational justification is available for it, the belief content is reliably self ascribed when it represents the participant's *current* state of mind. In other words, research participants should not believe that their relationship is doomed to failure only because they have been asked to think about extrinsic reasons to date their partners. But, if the research participants really believe that their relationship will fail, first-person authority is not compromised by their ascribing that belief to themselves.

How are we to judge what the participants' genuine state of mind is? It is a serious problem to arrive at identity conditions for beliefs and other attitudes. If I report an attitude and behave in a way that is compatible with it, there is good evidence that I have that attitude. The fact that I give reasons in support of that attitude is additional behavioral evidence. In the experiment by Seligman and colleagues, the participants endorse a belief that they are prepared to justify with reasons, e.g. that their relationship is not satisfactory and will not last. Even if we suspect that their other attitudes and future behavior will be at odds with their reported attitude (and with the prediction they make on the basis of it), we have no reason to believe that they are deceiving us (or themselves).

The dating couples study by Seligman et al. does not show that the attitudes reported by the research participants are unrepresentative of what they think or feel. Just because an attitude has been arrived at through manipulated evidence or biased reasoning, this does not mean that it will be fickle and that it will fail to be action guiding. But given the results of the further studies by Wilson et al. (1984), it is safe to project widespread failure in the predictions made on the basis of the attitudes reported by the participants

who engaged in reason giving and whose evidence was manipulated. It is likely, then, that participants will have been guilty of an attitude/behavior inconsistency, by acting in a way that is not compatible with their reported attitude.

Does the mismatch between reported attitudes and long-term behavioral dispositions undermine first-person authority? Inconsistencies between attitudes and behavior can be due to a variety of factors, such as a genuine and motivated attitude change or instability caused by an unresolved tension among conflicting attitudes. As Ferrero (2003) puts it, temporal stability of reported attitudes is not necessarily a precondition for first-person authority. Attitudes do not need to be persistent in order to be the object of knowledge. Let me sum up. Neither the rationality nor the stability of the reported attitudes should be a precondition for first-person authority. In order to maintain that research participants lack first-person authority over their reported attitudes, we would need to show that, at the moment of reporting their attitudes, they did not know what the content of their attitudes was or did not know that they had those attitudes. If participants reported something that they did not believe, but sincerely, then they would be mistaken about what they thought or felt, and they would experience a failure of first-person authority over their attitudes in spite of apparently meeting all the conditions for authorship. This would be a problem for a view such as Moran's, according to which authorship gives rise to first-person authority. But if they reported what they believed, it seems that, based on considerations about the rationality and stability of the reported attitudes, we have no reason to deny first-person authority to them.

3.2. Causal efficacy

One argument for the view that reason giving does contribute to first-person authority is that by weighing up evidence subjects are able to determine what their attitudes are. This claim is disputed in the interpretation of the psychological evidence. Giving reasons is presented as largely causally inert. When reason giving is a causally inert process, I *discover* rather than *determine* the content of the attitudes I report, even when I endorse them on the basis of reasons. The upshot is that there is no *agential* route to first-person authority: reason-searching is merely a heuristic and not a deliberative process, given that the content of the reported attitudes is established independent of the process of reason giving and is not affected by it. The distinction between the interpretive and the deliberative stance collapses and there is no need to distinguish between reasons contributing to the formation of an attitude and reasons offered in confabulation, because there are *no* reasons of the former type. I acquire beliefs about which beliefs I have in much the same way in which I acquire beliefs about what other people believe, and reason giving is epiphenomenal.

Philosophers stress the role of reason giving in deliberation whereas psychologists argue that reason giving is largely irrelevant to attitude formation. Moran emphasizes the active causal role of reason giving for the formation of attitudes, whereas Wilson opts for a deflationary account of reason giving as an ineffectual add-on. Who should we believe? Unexciting as it may sound, the truth lies somewhere in the middle. There are situations in which attitude formation occurs almost entirely below the level of conscious reflection, and in those situations psychologists are right that reason giving plays at best the role of advocacy. There are beliefs that just pop up in our belief box independent of a process of

deliberation, and whose causal history is not transparent to introspection (Wilson 2002, 97-98). A belief, say, has been formed, and one provides reasons for that belief that are aimed to show oneself and other agents how the belief makes sense within a certain narrative of the self. 'This is me, I don't like compromises'. 'It would have been silly to ignore her advice'. 'I never spend time with my friends; I could not have left the pub earlier'.

But even in the circumstances in which attitudes are the result of gut reactions, or subjects find themselves with attitudes they haven't explicitly formed by weighing up evidence, giving reasons is valuable, and there is a sense in which it is still an act of authorship. Subjects are not always epistemically blameworthy for not having access to the mechanisms responsible for the formation of the attitudes they find themselves as having. Reason giving cannot always contribute to uncover the causes of those attitudes, but, when the reasons offered are good reasons, it has other epistemic benefits: it enhances awareness of having those attitudes, and it is instrumental to creating connections between those attitudes and other attitudes subjects have, either allowing subjects to develop a coherent narrative or highlighting a clash that can give rise to the revision of their new or prior attitudes. Think about cognitive behavioral therapy as an example of how this process can be beneficial to the subject, even when the attitudes reported and defended with reasons are not 'optimal' or rational, and when it is not at all clear to the subject how they were formed.

Analyzing reasons brings about changes in the self: even the psychological evidence concedes this, although it suggests that those changes are usually for the worse. The point is that the attitudes endorsed as a consequence of the reason-giving exercise might be

significantly different from (and less desirable than) the ones reported at the beginning, but they are not necessarily less representative of the self. Even if the initial attitudes were not formed via deliberation, the opportunity for reason giving enables subjects to endorse the attitudes they find themselves as having on the basis of the reasons they take themselves to have. Whether the results of this reason-giving exercise are epistemically desirable depends on the details of each case. There are circumstances in which it might be more beneficial to be at the mercy of one's unconscious dispositions than to endorse attitudes that are based on non-rational processes of deliberation and justification (which I take to be the message of part of the introspection-effects literature). In other circumstances, reason giving will contribute to a greater awareness of one's attitudes, priorities and values, and will provide to subjects the resources to include their attitudes in a general narrative of themselves. This narrative, in turn, will play a causally active role in shaping explicit acts of deliberation and justification, by which the subject will be able to determine, and not just subscribe to, future attitudes (Velleman 2005).

3.3. Knowing why

The behavior of the dating couples in the experiment by Seligman and colleagues can be interpreted as follows. They knew what their attitudes were at the time when they reported them, no matter how representative of their future behavior those attitudes would have turned out to be. But they did not know *why* they had those attitudes. They did not realize that they formed those attitudes on the basis of partial evidence about their relationship, evidence that had been manipulated by the experimenter. Can we regard these research participants as authoritative with respect to the belief that their relationship

is not going to be long-lasting, if they ignored the reasons for their prediction? Described in these terms, the experimental results seem to be a good example of failure of first-person authority, because in this context a well-informed third person (e.g. the experimenter) could answer the why-question more reliably than the first person (e.g. the research participant).

But this move assumes very demanding conditions for knowledge of one's attitudes (again). If self knowledge must include not just knowledge of the content of one's conscious attitudes and knowledge of who is having those attitudes, but also knowledge of *why* the person having those attitudes has them, then self knowledge is made close to unattainable. In those cases in which there is no introspective access to the mechanisms responsible for forming or endorsing attitudes, self knowledge seems to require formal training in scientific psychology!

This implicit condition for first-person authority can be formulated in two ways:

> (a) In order to have knowledge of my having an attitude, *I cannot be mistaken about* the reason why I report and endorse that attitude.
>
> (b) In order to have knowledge of my having an attitude, *I need to be aware of* the reasons why I report and endorse that attitude.

Condition (a) sounds more plausible than condition (b). Condition (a) says that in order to have knowledge of my attitudes, I cannot have false beliefs about the reasons for my having those attitudes; no 'mistakes' are tolerated. Condition (b) says that in order to have knowledge of my attitudes, I must know what there is to know about the reasons for my having those attitudes; ignorance is not tolerated.

What could be the motivation for thinking that (a) and (b) are plausible conditions on first-person authority? One motivation could be that intentions and other attitudes are identified (also) on the basis of why subjects have them. Not knowing why I have an attitude, then, would mean not knowing *which* attitude I have. Let's the recall the example we discussed earlier. I intend to stay in the pub until closing time. Having true beliefs about why I have that intention, or being aware of the reasons for having it, may be necessary for me to know that I have that particular intention as opposed to another.

I might not be fully aware of the reason why I intend to stay in the pub until closing time, because I might not recognize the extent to which obtaining my friends' approval is important to me. When I endorse my intention to stay in the pub until closing time on the basis that I am having a good time, I am referring to a new attitude which happens to have a lot in common with the attitude caused by my unconscious desire to obtain my friends' approval. The content of the intention (*to stay in the pub*), and the commitment to that content expressed by the attitude (*intending* to stay in the pub) are the same, but the reasons for forming or endorsing the attitudes are different.

Let me summaries. If the reasons for forming or endorsing an attitude are relevant to establishing which attitude I report, ignoring why I have an intention compromises my authority over that intention. But, in the case at hand, it is open to the friend of authorship to argue that when we provide reasons for an attitude, we form a new attitude. Thus, there are really two attitudes, one acquired for reasons that are not transparent to introspection, and the other formed or endorsed on the basis of reasons. On this view, the attitude I endorse via reasons is not the one caused by the unconscious desire to obtain my friends'

approval, but it is the one motivated by the conscious desire to continue enjoying a pleasant evening.

What is the role of reason giving then? Through an analysis of my reasons, I don't gain any further insight into the reasons for my previous intention (the one caused by the desire to obtain my friends' approval), but I come to know that I have an intention that has the same content (the intention supported by my desire to continue enjoying the evening). Notice that when we describe the situation in these terms, my ignorance of the reasons for the initial intention to stay in the pub is not aggravated by reflection. Moreover, depending on whether we think that reason giving can be causally efficacious, there needs be no deception: I may be right that my desire to continue having a good time in the pub motivates my (further) intention to stay.

If the reasons for forming or endorsing the attitude do not matter to which attitude I report, the conditions for first-person authority formulated in (a) and (b) are not motivated, and first-person authority is neither enhanced nor undermined by reflection.

There are two psychological models that attempt to explain the instability of attitudes when subjects are asked to provide reasons for them, and they neatly map onto these two ways of describing the situation. The first account appeals to retrieval mechanisms and the idea that attitudes towards an object or person are constructed mainly on the basis of those aspects of the object or person to be evaluated that are easily accessible to the subjects when they attempt to retrieve the relevant information (Wilson et al. 1984). By looking for reasons for a previously reported attitude, subjects justify it on the basis of their most accessible reasons. These reasons might be different from what caused the attitude to be there in the first place.

According to the other model, 'attitudes are freshly computed based on available contextual cues' (Sengupta and Fitzsimons 2004, 711). Thus, attitudes are re-formed from scratch as a result of any new search for reasons. Both models explain the effects of evidence manipulation that subjects experienced in the dating couples study, but on the latter model participants do not have a definite set of attitudes that vary according to the retrieval circumstances. Rather, they construct their attitudes as they go along, using a variety of strategies that depend on the context of self ascription (Bettman et al. 1998). Research on preference reversals (Tversky and Thaler 1990, 210) also supports this view: attitudes, values and preferences are 'constructed in the process of elicitation'.

The former model assumes that the same attitude can be endorsed for different reasons. In the description of the situation according to which there is only one attitude, there is no real motivation for accepting such demanding conditions as (a) and (b) on first-person authority over reported attitudes. Moreover, the outcome of the reason-giving exercise is to allow the subject to justify her intention on the basis of *some* reasons rather than none, and this process might be beneficial by increasing the subject's awareness of having the attitude and promoting the inclusion of the attitude in a system in which it coheres or clashes with prior attitudes.

The latter model argues that attitudes are constantly re-formed if reasons in their support are being analyzed. In the description of the situation according to which there are really two attitudes, reason giving is not responsible for the ignorance of the reasons for the original attitude, but may be responsible for the formation or at least the genuine endorsement of the second attitude.

Conclusion

The significance of the agential route to self knowledge has been challenged on the basis of psychological data on the introspection effect and on elusive reasons. The interpretation of these data questions the role of introspective deliberation and justification for the acquisition of self knowledge. Even when the conditions for authorship are apparently met, as in the dating couples study, the attitudes reported as a consequence of searching for reasons are neither stable nor rational; and they do not seem to determine future behavior. Moreover, research participants ignore or are deceived about the reasons for their reported attitudes.

However, the argument against reason giving as a source of self knowledge works only if we accept very demanding conditions for first-person authority. I have shown that those conditions should be resisted: one can have first-person authority over an attitude even if the attitude is irrational and unstable; even if the reasons given for endorsing it are not the reasons why it was formed; and even if one is not aware, or is mistaken about, the reasons why one has formed or endorsed it.

Limited as it is, both in rationality and causal efficacy, authorship has a very important role to play. The psychological mechanisms that lead subjects to form an attitude can sometimes be partly or wholly unavailable to introspection and subjects often exercise no control over their behavioral dispositions. But by giving what they take to be their best reasons for the attitudes subjects find themselves as having, they gain control. At the end of the process of deliberation or justification, they might renew the commitment to the content of a previously reported attitude, but often on the basis of different reasons, or change their attitude and make a different commitment, which might meet independent

standards of rationality to a greater or lesser degree, depending on the quality of their reasoning.

These introspection effects should be expected consequences of analyzing reasons. Motivated attitude shifts are likely to increase responsibility and self awareness when the processes leading up to authorship are rational processes, e.g. when they take into account all the relevant available evidence and are based on the subject's best reasons. But even when the processes of deliberation and justification that give rise to new commitments are not optimally rational, the undesirability of these attitude shifts should be seen as a consequence of a breakdown of rationality and should not be seen as having negative implications for self knowledge. Reason giving has additional epistemic benefits: it increases awareness of having the reported attitudes and allows one to include them in a system of other attitudes, thereby constituting a first step towards the creation of a narrative of oneself that might demands coherence and motivate further attitude shifts.

Acknowledgements

I am grateful to Matteo Mameli, Matthew Broome, Jordi Fernández, and especially Edoardo Zamuner for extended comments on previous versions of this paper. I am also grateful to three anonymous reviewers for constructive criticism. Audiences at departmental seminars or workshops at the University of Reading, the University of Glasgow, the University of Adelaide, Macquarie University, the Australian National University and the University of Northampton provided challenging discussion and helpful feedback.

References:

Bettman J., Luce M., Payne J. (1998). Constructive consumer choice processes. *Journal of Consumer Research*, *25*, 187-217.

Bortolotti L. (2004). Intentionality without Rationality. *Proceedings of the Aristotelian Society*, CV (3), 385-392.

Bortolotti L. (2005). Can we Interpret Irrational Behavior? *Behavior & Philosophy,* 32(2), 359-375.

Bortolotti L. and Broome, M. (forthcoming). Delusional beliefs and reason-giving. *Philosophical Psychology*.

Burge T. (1988). Individualism and self-knowledge. *Journal of Philosophy, 85*, 649-663.

Campbell J. (1999). Schizophrenia, the space of reasons and thinking as a motor process. *The Monist, 82(4),* 609-625.

Carman T. (2003). First persons: On Richard Moran's *Authority and Estrangement*. *Inquiry, 46,* 395-408.

Clarke S. (forthcoming). SIM and the City: Rationalism in Psychology and Philosophy and Haidt's Account of Moral Judgment. *Philosophical Psychology*.

Dennett D. (1989). The origins of selves. *Cogito, 3,* 163-73.

Dennett D. (1992). The self as a center of narrative gravity. In F. Kessel, P. Cole and D. Johnson (eds.) *Self and Consciousness: Multiple Perspectives*, Hillsdale, NJ: Erlbaum.

Dretske F. (2006). Minimal rationality. In S. Hurley and M. Nudds (eds.), *Rational Animals?* Oxford University Press.

Evans G. (1982). *The Varieties of Reference* (ed. by J. McDowell). Oxford University Press.

Fernández J. (2003). Privileged access naturalized, *Philosophical Quarterly*, *53 (212),* 352-372.

Ferrero L. (2003). An elusive challenge to the authorship account. *Philosophical Psychology, 16(4)*, 565-567.

Gallagher, S. (2007). Sense of agency and higher-order cognition: Levels of explanation for schizophrenia. *Cognitive Semiotics, 0*: 32-48

Gerrans P. (2001). Authorship and ownership of thoughts. *Philosophy, Psychiatry and Psychology, 8 (2-3)*, 231-237.

Haidt J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*, 814-834.

Halberstadt, J. B., and Levine, G. L. (1999). Effects of reasons analysis on the accuracy of predicting basketball games. *Journal of Applied Social Psychology*, 29, 517-530.

Heal J. (1988). Understanding other minds from the inside. In A. O'Hear (ed.) *Current Issues in Philosophy of Mind*. Cambridge University Press, 83-100.

Heal J. (2001). On first-person authority. *Proceedings of the Aristotelian Society, 102 (1)*, 1-19.

Heil J (1988). Privileged access. *Mind, 47*, 238-235.

Hixon G. and Swann W. (1993). When does introspection bear fruit? Self-reflection, self-insight, and interpersonal choices. *Journal of Personality and Social Psychology, 64*, 35-43.

Hoerl C. (2001). On thought insertion. *Philosophy, Psychiatry and Psychology, 8(2-3),* 189-200.

Holt L. (1993). Rationality is hard work: an alternative interpretation of the disruptive effects of thinking about reasons. *Philosophical Psychology, 6(3),* 251-266.

Kahneman D., Slovic P., and Tversky, A. (Eds.). (1982). *Judgements under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Lawlor K. (2003). Elusive reasons: a problem for first-person authority. *Philosophical Psychology, 16 (4)*, 549-564.

Moran R. (2001). *Authority and Estrangement: an Essay on Self-knowledge*. Princeton University Press.

Moran R. (2004a). Précis of Authority and Estrangement. *Philosophy and Phenomenological Research, LXIX (2),* 423-426.

Moran R. (2004b). Replies to Heal, Reginster, Wilson and Lear. *Philosophy and Phenomenological Research, LXIX (2),* 455-472.

Mullins S. and Spence S. (2003). Re-examining thought insertion. *British Journal of Psychiatry, 182*, 293-298.

Nisbett R. and Wilson T. (1977). Telling more than we can know: verbal reports on mental processes. *Psychological Review, 84,* 231-259.

O'Brien L. (2003). Moran on agency and self-knowledge. *European Journal of Philosophy, 11,* 375-390.

Peacocke C. (1998). Conscious attitudes, attention, and self-knowledge. In C. Wright, B. Smith, and C. Macdonald (eds.) *Knowing our Own Minds*. Oxford: Clarendon Press, 63-98.

Pettit P. and Smith M. (1990). Backgrounding desire. *The Philosophical Review, 99(4)*, 565-592.

Ryle G. (1949). The Concept of Mind. University of Chicago Press.

Seligman C., Fazio R. and Zanna M. (1980). Effects of salience of extrinsic rewards on liking and loving. *Journal Personality and Social Psychology, 38*, 453-460.

Sengupta J. and Fitzsimons G. (2004). The effect of analyzing reasons on the stability of brand attitudes: a reconciliation of opposing predictions. *Journal of Consumer Research, 31,* 705-711.

Sims A. (2003), *Symptoms in the mind* (third edition). Saunders.

Stanovich K. E. (1999). *Who is Rational? Studies of Individual Differences in Reasoning*. Erlbaum Associates.

Stephens G. and Graham G. (2000). *When Self-Consciousness Breaks: Alien Voices and Inserted Thoughts*. MIT Press.

Tordesillas R. (1999). Thinking too much or too little? *Personality and Social Psychology Bulletin*, *25*(5), 625-631.

Tversky A. and Thaler R. (1990). Anomalies: Preference Reversals. *Journal of Economic Perspectives*, 4(2), 201-211.

Velleman J. (2005). The self as narrator. In J. Christman and J. Anderson (eds.), *Autonomy and the Challenges to Liberalism: New Essays*. New York: Cambridge University Press, 56-76.

Wegner D. (2002). *The illusion of conscious will*. MIT Press.

Wilson T. (2002). *Strangers to ourselves*. Harvard University Press.

Wilson T. and Dunn E. (2004). Self-knowledge: its limits, value, and potential for improvement. *Annual Review of Psychology, 55*, 493-518.

Wilson T. and Hodges S. (2004). Effects of analyzing reasons on attitude change: the moderating role of attitude accessibility. *Social Cognition 11*, 353-366.

Wilson T., Hodges S. and LaFleur, S. (1995). Effects of introspecting about reasons: inferring attitudes from accessible thoughts. *Journal of Personality and Social Psychology, 69,* 16-28.

Wilson, T. D., Hodges, S. D., & LaFleur, S. J. (1984). Effects of Analyzing Reasons on Attitude-Behavior Consistency. *Journal of Personality and Social Psychology*, 47(1).

Wilson, T. D., & Kraft, D. (1993). Why do I love thee? Effects of repeated introspections about a dating relationship on attitudes toward the relationship. *Personality and Social Psychology Bulletin*, *19*, 409-418.

Wilson, T. D., Kraft, D. & Dunn, D. S. (1989). The Disruptive Effects of Explaining Attitudes: The Moderating Effect of Knowledge about the Attitude Object. *Journal of Experimental Social Psychology*, *25*, 379-400.

Wilson T. and Schooler J. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology, 60*, 181-192.

Wilson T., Dunn D., Bybee J., Hyman D. and Rotondo J. (1984). Effects of analyzing reasons on attitude-behavior consistency. *Journal of Personality and Social Psychology, 47*, 5-16.

Wright C. (1998). Self-knowledge: the Wittgensteinian Legacy. In C. Wright, B. Smith and C. Macdonald (Eds.). *Knowing Our Own Minds*. Oxford: Clarendon Press, 15-45.

[1] By rationality in this paper I mean 'conformity to norms of correct reasoning'. I am aware that this is not the only notion of rationality relevant to this context. I discuss merits and limits of different notions of rationality as applied to intentional states elsewhere (Bortolotti 2004 and 2005).

[2] In the paper I am interested in the ownership and authorship of a variety of attitudes (preferences, beliefs, motivated desires, decisions, intentions etc.) but will refer primarily to the case of beliefs in my examples.

[3] For a similar account of self knowledge, where one comes to know what one believes by considering the reasons for what one believes, by looking *outward* rather than *inward*, see Fernández (2003).

[4] For different accounts of the capacities impaired in thought insertion, see Gallagher (2007); Mullins and Spence (2003); Hoerl (2002); Campbell (1999).

[5] For a discussion of the evidence see Clarke (forthcoming).

[6] There are studies on preferences for jams which have been used to illustrate the effect of reason giving on attitude change. For instance, see Wilson and Schooler's 'Jam Taste Test' (1991). In the study, people express preferences that are aligned with those of experts when they do not think about reasons for their attitudes.

[7] There is nothing incoherent about having irrational attitudes. I argued elsewhere that irrationality does not compromise the ascription of intentional states or the explanation and prediction of behaviour in intentional terms (Bortolotti 2004 and 2005).

[8] For the purposes of the present discussion I will not distinguish the effects of *spontaneous* reason giving (when the subject endorses the content of a belief via reason giving without being asked to do so) from those of *prompted* reason giving (when the subject is asked to provide reasons for a reported belief).

[9] Notice that this premise does not seem to be justified on the basis of Moran's account of authorship. Moran suggests that reason giving provides an additional route to first-person authority (different from privileged epistemic access), and thus does not claim that epistemic and agential authority are cumulative.

[10] See Tiberius (2008) for a detailed analysis of these studies and of their implications for the relationship between reflection and well-being.