DOI: 10.1111/j.1468-0378.2007.00266.x

The Moral Development of First-Person Authority

Victoria McGeer

> For any detailed description of the complexity of human nature, of the insurgence of instinct in the garb of reason, of the multifarious play of the social environment in the individual ego, and of the individual ego on the social environment, I had to turn to the novelists and the poets.
> (Beatrice Webb, *My Apprenticeship*)

It is a natural, commonsense assumption that human beings who are competent in their understanding and use of folk-psychological concepts (e.g. 'belief', 'desire', 'intention', 'fear', 'hope', 'jealousy' and the like) have a special kind of authority with respect to claims they make about their own minds, in particular about their own intentional attitudes. One way to capture this special sense of authority is to argue that such claims are subject to a 'default hypothesis' of correctness (e.g. Wright 1991: 143–4). If I claim to be *upset* or *happy* about something or to have a *yearning* for plum pudding, then, all things being equal (i.e. assuming I am sane, and sincere, and not deeply distracted), the appropriate default presumption is that such claims are true. This presumption must be carefully understood, of course. On the one hand, it does not amount to endorsing a person's infallibility or even incorrigibility with respect to the claims they make about their own minds; others may successfully challenge them from the point of view of making sense of the overall pattern of that person's behaviour. On the other hand, the evidence of someone's 'unextorted word' about what they think, desire, or feel takes a lot to defeat (Dennett 1987: 20). For not only must the rest of the person's behaviour speak strongly against taking them at their word; there must be some reasonable account of how they have failed to maintain first-person authority in the particular case. In other words, the idea of a special kind of authority attaching to first-person claims brings with it the demand for special explanations in the case of failure.

There is a long tradition in philosophy of attempting to explain the special qualities of first-person authority in terms of a privileged epistemological relation a person bears to her own mind. Yet this purely epistemological approach not only encounters serious difficulties in its own terms; it fails to account for a critical feature of first-person authority even if these difficulties could be overcome. It fails to explain the close connection between acknowledging someone's authority over her own psychological states and treating her as the sort of agent who can be held responsible for what she thinks and does. Such an

European Journal of Philosophy 16:1 ISSN 0966-8373 pp. 81–108 © 2007 The Author. Journal compilation © Blackwell Publishing Ltd. 2007, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

agent is 'psychologically capable', as I will say, in a distinctive kind of way. Anyone who is lacking in first-person authority is going to be lacking in this psychological capability: she will be lacking the capacity to think and operate as a rational, responsible, self-directed human being. Thus, in the normal case, we not only defer to others' claims about themselves, we *owe* them that deference so far as we treat them as rational, responsible agents. In order to capture this aspect of first-person authority, we need to abandon the epistemological model in favour of what I have elsewhere called the 'agency model of authoritative self-knowledge' (McGeer 1996; see also, Moran 2001). This model traces first-person privilege not to an agent's capacity for epistemic accuracy in self-ascription, but to a capacity to shape or determine her own states of mind. The agent has a privileged authority in self-ascribing intentional states because it is she who makes it the case that she deserves to be ascribed these states; she has 'maker's knowledge', not the knowledge of a particularly accurate perceiver or detector.

This paper is an attempt to connect the agency model of self-knowledge with broader issues in moral psychology, particularly the psychology of moral development. In section 1, I briefly present the epistemic model in order to highlight certain structural features of it that disable it from giving any obvious or ready explanation of the intimate connection between first-person authority and an agent's psychological capability. In section 2, I advocate replacing the epistemic model with an 'agency' alternative, and distinguish my preferred version of this model from an importantly influential variant elaborated by Richard Moran (Moran 2001, 1997). In section 3, I argue that, despite its many virtues, Moran's version has unattractive implications for assessing rational agency and moral performance. Finally, in section 4, I try to show that Moran's version has these unattractive features because the ideal of rational agency he articulates is fundamentally unsuited for creatures like us, in particular creatures with our sort of (moral) developmental history. In keeping with Moran's own practice of using literary examples to clarifying effect, I rely on a few 'case studies' drawn from George Eliot's novel *Middlemarch* to support the arguments of the last two sections.

1. Conceptual Limitations of the Epistemic Model of Authoritative Self-Knowledge

The attempt to explain first-person authority as a purely epistemological phenomenon begins straight-forwardly by taking the furniture of my mind, like the furniture of the world, as given. That is to say, my mind is presumed to be constituted by a rich array of psychological states and processes, which of course change over time under changing conditions. However, within this epistemological framework, the problem of first-person authority is not to explain how these states and processes come and go; it is rather to explain how I come to know about whichever ones are there. In particular, it is the problem of explaining how the way in which I come to know about my own states and processes, as against the way that others might come to know about them, gives me sufficient epistemic superiority to account for the authority normally vested in the judgements I make about them and sometimes articulate in self-report.

One obvious and commonsensical way of viewing the asymmetry between first- and third-person ways of knowing makes appeal to a privileged access I have to my own states that no one else can share. States of my own mind are presented 'immediately' to me, whereas others must infer their existence from their presumed causal role in mediating my interactions with the world. It has been standard to construe this immediacy on a quasi-perceptual model: I have an internal way of sensing my psychological states, just as I have an internal way of sensing the physical position of my own body, a faculty of introspection akin to the faculty of proprioception. This faculty of introspection need not be construed as anything so crude as an inner eye. These days philosophers tend to speak in terms of a reliable sub-cognitive tracking mechanism that provides a causal link between my (first-order) psychological states and processes and the second-order states that constitute my knowledge of them (Armstrong 1968, 1993). Still, the model can be viewed as broadly perceptual since my first-order states and processes are held to be independent existents that are only contingently related to my second-order beliefs about them (Shoemaker 1996: essays 10 & 11). Thus, the idea of immediacy is reconciled with the epistemological desideratum that self-knowledge constitute a genuine cognitive achievement. Of course, the sense of achievement here is rather formal and thin: self-knowledge is not *my* achievement in any real agential sense, except, perhaps, so far as I exert attentional control over the sub-cognitive mechanism that does the work. But this is just what we should expect from a broadly perceptual model. What I perceive, whether out in the world or in my own mind, depends on where I direct my attention.

The conceptual simplicity of this model has enormous appeal. Yet its focus on explaining our putative epistemic superiority with respect to our own intentional states, long considered the fundamental problem of first-person authority, distracts attention from another pressing problem that any fully adequate account of this phenomenon should be able to address—namely, why having first-person authority is so intimately connected with a capacity for rational, responsible self-directed agency; and correspondingly why having such authority is *not* granted just by virtue of having (good) evidence about our own minds. If the epistemic model lacks the explanatory resources to account for this connection, then we have good reason to reject it.

Before analysing this problem in more detail, there is one methodological concern that ought to be addressed. It may be argued that since the epistemic model does not aim to explain the connection between first-person authority and psychological capability, it should not be assessed according to this criterion; rather it should be judged solely on the basis of how well it succeeds at explaining and justifying the assumption of epistemic superiority.[1] In response, I have two observations. The first is to concede that a full critique of the epistemic model should certainly consider how well it fares at meeting its own explanatory desiderata. My own view, argued at length elsewhere, is that it does not fare very well, and that this in itself gives us very good reason to pursue the agency alternative discussed in section 2 below (McGeer 1996).[2] But my second observation, here more to the point, is that the philosophical treatment of any phenomenon must surely give equal attention to *explanandum* and *explanans*, developing through inquiry a fully satisfying account of the phenomenon to be explained as much as a fully satisfying explanation of it. For instance, a philosophical treatment of the problem of free will concerns itself not just with providing candidate explanations of this phenomenon, but also with probing the complexity of our commonsense notion in order to clarify just what kind of phenomenon it is that requires explanation. Hence, I do not think it is out of court to fault the epistemic model with failing to fully characterize, and so adequately explain, the phenomenon of first-person authority. We may be phenomenologically and/or traditionally primed to think of first-person authority as resting on a straightforward kind of epistemic superiority; but the proposal here is that there is more packed into the commonsense notion of this phenomenon than such a construal allows.

So what is the basis for claiming that an agent's psychological capability is conceptually linked to the phenomenon of first-person authority? To see this, consider again the issue of immediacy. Most philosophers agree that first-person authority attaches exclusively to psychological claims that are 'immediate' in the following sense: they are *not* based on evidence; in particular, they are not based on evidence from our own behaviour. Of course, we can make claims about ourselves based on observations of own behaviour—or, more likely, based on what others point out to us about our own behaviour: e.g. 'I guess I *am* a bit angry with her, given what I just did'. But what is peculiar about such claims is not just that they lose their authoritative status by assuming the profile of third-person claims; they also attest to a sense in which we are *alienated* from ourselves, at least with respect to the psychological states so claimed (Moran 2001). Now it is certainly true that there is an epistemological aspect to this alienation: it is a condition marked by the fact that the 'first-person' claims made in this mode are on the same epistemic footing as others' claims about us: they both arise from the evidence of our behaviour and must answer to that evidence in just the same sort of way. However, the language of alienation also indicates that we think something is psychologically amiss with a person who cannot know her own anger, say, in the immediate way that characterizes first-person authority. It smacks of a kind of failure of agency, even of responsibility. So with this contrast class in mind, we can see that the phenomenon of first-person authority raises two puzzles that must be solved together. The first is the familiar, seemingly epistemological one: how do we account for the (non-evidentiary) immediacy of certain first-person claims in a way that justifies their special authoritative status? But the second and connected puzzle is properly speaking a moral-psychological one: how do we account for the fact that knowing our own minds with the immediacy of first-person authority is bound up with the kind of psychological capability that speaks to an ideal of well-functioning agency, whereas knowing our own minds in any other way testifies to a certain sort of agential dysfunction?

Richard Moran, whose work has brought such clarity to this feature of first-person authority, makes the following compelling observation (Moran 2001: 90–4). Standard epistemological accounts that focus on solving the first of these puzzles in perceptual or quasi-perceptual terms simply haven't got the resources to give us a satisfying solution to the second. We can see this, he argues, by considering an epistemically idealised individual whose privileged inner access to the psychological contents of her own mind is such that her judgements suffer none of the errors that psychologists insist plague ordinary people. She has complete information, complete accuracy and complete reliability, making her an unquestionable epistemic authority on her own psychological states and processes. Yet, for all that, there may be something deeply impaired in her own powers of agency, for there is nothing in the idea of her having this super-perceptual capacity that gives her, by that token, any capacity to affect what comes and goes in her mind. Her spectatorial capacity may be entirely passive, making her position vis-à-vis her own mind no better or worse than another person's might be, could we endow that other person with telepathic powers. In neither case does such epistemic immediacy give the observing subject any authorial control over the mind being observed, making her relationship to that mind essentially third-personal. But while no agential dysfunction is implied in the telepath's case, there is something deeply wrong with a person thus related to her own mind. Whatever drum she is marching to, it is not of her beating. The credit (or blame) must be placed elsewhere, and she is stuck willy-nilly with living through the consequences.[3]

The message of this thought-experiment is clear. If we're to solve the moral-psychological puzzle, we need to make a deeper connection between agency and self-knowledge than the purely epistemological approach allows. In essence, we need to widen the scope of our explanatory framework in order to understand the ways in which our own powers of agency are involved in creating and maintaining the phenomenon of first-person authority.

2. The Agency Model of Authoritative Self-Knowledge

The reflections of the previous section point towards exploring an alternative model of first-person authority: what I have called the 'agency' model. The chief characteristic of this model is that it aims to explain whatever epistemic privilege we have in our first-person psychological claims in terms of a prior notion of agential privilege that somehow exploits the idea of authorship. In this section, I want to describe some of the features of the agency model by comparing two versions of it. One is the version I prefer (sketched in McGeer 1996); the other is Richard Moran's (Moran 2001). I focus on these two accounts for simplicity's sake.[4] In my view, they share much in common. But, as I will go on to show, there are differences between them that turn out to be particularly revealing for exploring the kind of psychological capability that underlies first-person-authority and thus establishes a normative ideal for well-functioning agency—i.e. the kind of agency that is admirably rational, self-directed and morally able.

To begin with some common ground, both versions of the agency model here considered regard the epistemic approach as resting on a false presupposition—viz., that the problem of first-person authority is not to explain how my (first-order) states and processes come and go, but rather how I come to know (authoritatively) about whichever ones are there. As we have seen in Section 1, the epistemic approach would put me in an essentially spectatorial, third-personal relationship with the contents of my own mind, no matter how privileged my observer's point of view is claimed by its proponents to be. Interestingly, this picture is often faulted for the 'metaphysically extravagant' way in which it portrays first-person privilege. But, as Moran insightfully remarks, the deeper problem is that 'under the *guise* of metaphysical extravagance this picture of privacy presents an essentially superficial view of the differences between my relation to myself and my possible relation to others. For in essence what we have here is a picture of self-knowledge as a kind of mind-reading as applied to oneself . . . ' (Moran 2001: 91). This picture promotes a 'reporter-predictor model' of authoritative self-knowledge (McGeer 1996).

Moving beyond this critical point, both accounts agree in a general way on how to amend the epistemic approach—namely, by showing how our own consciously directed agency is actively involved in making and maintaining the psychological states and processes we claim for ourselves. The idea is that our first-person authority rests not on any kind of observational privilege, but rather more directly on the authorial privilege of agency, a privilege that brings with it its own set of responsibilities. Some of these responsibilities are intra-personal—for instance, the responsibilities that go with being deliberatively rational; and others are inter-personal—for instance, the responsibilities that go with conducting ourselves coherently in our interactions with others.

The version of the agency model that I prefer stresses the inter-personal, social aspect of responsible agency in giving an account of authoritative self-knowledge. The rationale for this emphasis derives from the commonsense observation that success in our day to day social interactions depends on our being able to give explanations and make predictions of one another's behaviour from what Dennett has called the 'intentional stance'—the stance that involves the attribution of a rich profile of (rationalising) folk-psychological states and processes (Dennett 1987). Of course, we get an enormous leg up in adopting this stance so far as we are able to rely on what people claim about their own psychological states. Thus, it will be no surprise to find a normative dimension built into our practice of taking one another at our word. This normative dimension consists in a practical recognition that mutual understanding depends, not just on taking one another's self-attributions seriously, but even more significantly, on holding one another responsible to those attributions, and so learning to hold ourselves responsible to them as well. Hence, in order to be able players in the language-game of intentional attribution—i.e. in order to engage in ordinary social interaction—we must be willing to bring our words into line with our deeds and our deeds into line with our words. We must be committed to creating a profile of ourselves as agents who can be depended on to manifest the qualities of first-person authority, regulating our performance in accord with the claims we make about ourselves.

But how is this form of self-regulation actually achieved? Obviously, it depends on our learning, throughout development and even into adulthood, how to characterize our own attitudes, thoughts and feelings in folk-psychological terms. That is, it depends on our learning shared standards of what counts as having various kinds of propositional attitudes (beliefs, desires, wishes, fantasies, hopes, whims and so on), as well as what counts as having various kinds of emotions (anger, jealousy, delight, pride, *Schadenfreude*, and the rest). But more deeply than this, it depends on learning to take responsibility for our various first-order states, by actively involving ourselves in forming, reviewing, revising, suppressing, and selectively acting on them, making them into first-order states we authoritatively know about because we are the ones directly involved in generating and sustaining them.[5] Thus, when we are asked to report what's in or on our minds, we do not track or map internal mental states, as we might perceptually track or map objects and events in the world. Rather, we first reflect on how, given the current state of the world, we ought to be minded; and then we actively commit ourselves to displaying the mind—for example, the intentional attitudes—that we think appropriate. Thus, our first-person authority is dynamically sustained, not through a series of snapshot judgements that accurately capture some underlying independent phenomena, but rather as a matter of reflective attitudinal and behavioural co-ordination.

If this picture is correct, we can make better sense of the way in which others can challenge what we say about ourselves based on mismatches in an overall profile of what we say and do without yet challenging our first-person authority. For we manifest our authority—i.e. our capacity to regulate and so author our own minds—not just by getting ourselves right in a purely judgemental sense, but by being able to bring our words and deeds into comprehensible alignment according to the norms of folk-psychology. If we are faced with legitimate challenges to the claims we make about ourselves, it is always an option to bring things back into alignment by means of various restorative actions including retractions, apologies, explanations, excuses and behavioural adjustments. Knowing when such restorative actions are called for, and which are appropriate to make, is part of what it means to be an authoritative folk-psychological agent—an agent who is normatively regulated by the social demands of first-person authority.

We now can see how these inter-personal responsibilities of authoritative agency tie in with the more intra-personal, deliberative responsibilities that are particularly high-lighted by Moran in his own version of the agency model. For in this social agential mode, we are not just making our psychological self-ascriptions at the behest of random internal promptings; nor are we doing so in a way that joins with others in making empirical conjectures about what we actually think and feel. Instead, we are actively deliberating on what we ought to think and feel, producing states that we can *endorse*, that we can *avow*, that we can feel comfortable *committing* to as our own. We are *making up* our minds, to use Moran's apt phase—and I would add, governing ourselves accordingly.

Now, this way of putting things may sound objectionably voluntaristic.[6] But, as Moran insists, the impression is misleading. It is not that we are free to pick and choose whatever psychological states suit us best. It is rather that we engage our reason to determine what is appropriate to think, desire, and feel given how we find the world and our situation in it; and, in Moran's paradigm case, by so engaging our reason, our deliberative conclusions become the states that are active in us. So, for instance, if I ask: 'what should I believe under these circumstances?' then, again in Moran's paradigm case, the answer I come to as a result of my deliberations *is* what I now believe. The belief is made active in me through my deliberatively generated avowal of it. Moran calls this the paradigm case of authoritative agency because, in his view, 'there would be nothing that counted as agency or deliberation at all if a person could not generally claim the conclusion of his reasoning as making it the case that, as a matter of psychological fact, *this* is his belief about the matter' (Moran 2001: 120).

So where does the general thrust of this agency model leave us? In a much stronger position, I think, to rebut any naturalist complaint that first-person authority cannot be defended if it's discovered that we do not always judge best which states and processes are psychologically active in us and have a causal impact on our behaviour. For, once again, this sort of accuracy is not really to the point. As Moran nicely summarizes, 'the primary thought gaining expression in the idea of "first-person authority" may not be that the person himself must always "know best" what he thinks about something, but rather that it is *his business* what he thinks about something, that it is up to him. In declaring his belief, he does not express himself as an expert witness to a realm of psychological facts, so much as he expresses his rational authority over that realm' (Moran 2001: 123–4). This way of understanding first-person authority presumes that the agent can exert a reasonable amount of deliberative control over the contents of her mind; that her deliberations can make a difference to what she thinks and feels; that they are not just epiphenomenal fluff (cf. Pettit and Smith 1996).

At this point, however, a question arises that generates a divide between the two versions of the agency model discussed so far. The question is whether the deliberative component is sufficient on its own to ensure the sort of rational authority that is characteristic of a psychologically well-formed agent, or whether it needs to be supplemented by a (non-deliberative) regulative component. The psychologically capable agent asks in deliberation what she ought to believe or desire or intend and forms a conclusion. Does it follow that she spontaneously believes or desires or intends as she concludes she should? Or is a continuing exercise of self-regulation and self-control sometimes necessary in order to ensure that she stays faithful to that conclusion?

Moran himself is a deliberative purist. On his view, being 'rationally autonomous', and thereby authoritative, with respect to our psychological states and processes is to be distinguished from exercising 'rational control' over those states and processes. As he says:

> The specifically first-personal responsibility that a person has for his own desire is essentially not instrumental. The person's responsibility here is to make his desire answerable to and adjustable in light of his sense of some good to pursue. It is not a responsibility that reduces to the ability to exert influence over one's desires, and that is why the idiom of 'control' is misleading in this context. When the desire is (already) the expression of the person's reasons, there is no *need* for exerting any control over it. As in the case of ordinary theoretical reasoning, which issues in a belief, there is no further thing the person *does* in order to acquire the relevant belief once his reason has led him to it. (Moran 2001: 118–9)

I don't think this is right. First-person authority can very much depend on our abilities to exercise rational control—though I would prefer to speak of 'self-regulation' or 'self-governance'. First-person judgements—judgements we make about what *to* believe or desire—have a certain 'commissive quality': they are judgements made in the indicative mode—I *do* believe this—that commit us to speak and act in ways commensurate with those judgements. However, in order to follow through on such judgements and so manifest the qualities of agents with first-person authority, we sometimes need to regulate our thoughts and deeds in ways that do not fully accord with Moran's notion of autonomy. To flesh out the precise nature of this disagreement, I begin by characterizing two cross-cutting dimensions in accord with which varieties of self-regulation can be distinguished: one is the relative *automaticity* with which they take place, and the other is their relative *instrumentality.* I consider each of these dimensions in turn.

Much self-regulation we do 'automatically'—i.e. without much explicit attention. Our training in the normative discipline of folk-psychology develops in us self-governing habits of mind whereby we often make use of our own intentional self-ascriptions as resolutions or reminders, to help instil or reinforce tendencies and inclinations that (normatively) fit with these ascriptions, and to disempower tendencies and inclinations that do not (McGeer 1996: 510). In addition, we may develop idiosyncratic habits of thought or action to circumvent what we know to be specific cognitive or affective weaknesses. For instance, we may avoid dwelling imaginatively on certain topics; we may gravitate towards certain kinds of friends or work environments; we may establish a variety of social or physical constraints and incentives; we may even adopt particular meditational or other therapeutic practices. Some of these habits we may develop without any explicit awareness that we are managing our cognitive/affective weaknesses. Further, they may become so ingrained that it takes very little effort or thought to continue in them. However, there may also be circumstances in which our self-managing techniques need to be explicit, effortful and well-planned if we are to cleave to the attitudes we have deliberatively endorsed

(McGeer and Pettit 2002: 288–90). For instance, a beginner pilot may know, both theoretically and by dint of explicit drilling, that proprioceptive cues can be deeply misleading when flying under certain conditions. However, she may also find it extremely difficult to resist their 'seat of the pants' allure. Her instruments deliver one message, her gut another—and though she may well know what she *ought* to do (follow the instruments), she may also know that stress or other forms of cognitive load can weaken the grip her deliberatively endorsed judgements have on her. Hence, it pays to have set in place certain 'fail safe' devices on which she can rely—for instance, warning buzzers or an attentive co-pilot or air-traffic controller telling her to mind her instruments if she seems to be going off track.

This last example helps brings to the fore a second dimension along which we can distinguish practices of self-regulation—viz., their *instrumentality.* Here is where I find myself in sharpest disagreement with Moran's view, although the nature of this disagreement must be carefully specified. For instance, Moran claims 'it would be an expression of the failure of reasoning were it to terminate not in conviction itself but rather in my apprehending a particular thought, or even appraising it as best, which I then need to find some *way* to make my own, my actual state of mind' (Moran 2001: 131). At the level of judgement, this is surely correct. An agent's first-personal deliberation aims at answering the question: what *shall* I believe, or want, or intend? Hence, its objective is to endorse the *contents* of particular psychological states; to arrive—as we might say—at a certain kind of judgement: *This* is what I shall believe, or desire, or intend. My deliberations end in conviction; and in so far as I am a well-functioning rational agent, this judgement becomes, through my deliberations, psychologically active in me. Still, as Moran might well agree, having the judgement become my psychologically active state of mind is rather different from having a full-blown belief or desire become my actual 'state of mind'. For beliefs and desires are complex dispositions to think, speak, feel and otherwise operate in various mental and physical ways. To endorse the claim that p is one thing; to fully instantiate the state of believing that p is another. Thus, one's deliberative judgement that p—or that the evidence supports p, or whatever—may leave further work to be done in order to ensure that one counts as being a fully robust believer that p.[7]

But what sort of work might this involve? On Moran's deliberatively purist view, it might involve reflecting on what thoughts or behaviour are (or are not) commensurate with believing that p, reminding oneself of the good reasons one has for believing that p, or even enlisting the argumentative support of others to bolster one's understanding of why p is the thing to believe. Such deliberative practices, though often effortful, are not by that token inimical to an agent's rational autonomy. But the same cannot be said for other self-regulative practices, however much they may be guided by reason. As Moran explains:

> In various cases a person may produce in himself various desires, beliefs, or emotional responses, either by training, mental discipline, drugs, the cooperation of friends, or simply by hurling himself into a situation that

> will force a certain response from him. But exercising this sort of control over one's attitudes is not the expression of 'activity' relevant to autonomy or rational authority. In such cases of producing a desire in oneself, the resulting attitude is still one I am essentially passive with respect to. It is inflicted on me, even if I am the one inflicting it. (Moran 2001: 117)

On the face of it, these various techniques of self-regulation seem a little strange to lump together. The kind of targeted rational control we're able to exercise through training, mental discipline or even the careful selection of our environment seems rather different from the kind of broad-stroke psychological effects we might be able to induce through taking drugs. However, in Moran's view, these techniques all count as *external* manipulations because they are not constitutive of deliberative processes in and of themselves. Consequently, any individual who needs to resort to these is, by that very necessity, 'alienated' from himself with respect to the states so manipulated. Furthermore, according to Moran, if he is so alienated, then there is something psychologically amiss with the way he is functioning as a rational being. For now he is limited in how far he can adopt the stance from which he is able to 'declare the authority of reason over his beliefs and his actions' (Moran 2001: 127). Thus, in a traditional Kantian play on the idea of rational autonomy, Moran insists that a person's rational activity is compromised to the extent that his psychological states and processes do not 'spontaneously' conform—i.e. through deliberation—to the dictates of reason. That is to say, his rational activity is compromised to the extent that he uses any extra-deliberative means to instantiate the dispositional profile his reason exhorts. For now this activity cannot consist in pure deliberation about what *to* believe, desire or intend, but must involve itself instead some messy empirical consideration of how best to bring about *causally* what ought to arise spontaneously as the expressive outcome of deliberation itself (Moran 2001: 117–8).[8] The agent's rational activity is thus uncomfortably divided between what Moran describes as two fundamentally different perspectives: the first, which he calls the 'transcendental perspective of agency', is irreducibly first-personal, the maker's perspective, the perspective from which the agent asks, 'what shall I—*qua* agent—believe?'; and the second, the 'empirical perspective of psychological facticity', is irreducibly third-personal, an observer's perspective, the perspective from which the agent stands back from herself and poses the question, 'what do I—*qua* subject—believe?'

3. Two Difficulties With Moran's Deliberative Purism As A Condition On Well-Formed Agency

In the last two sections, we've aimed at developing an account of authoritative self-knowledge that does two things: (1) it gives a substantial psycho-philosophical explanation of the privileged relation an agent has over her own psychological states and processes sufficient to justify the assumption of first-person authority; and (2) it makes clear why having this privileged relation matters for manifesting the kind of psychological capability that is the hallmark of well-functioning authoritative agency. The agency model was put forward as satisfying both explanatory desiderata. It cashes out the notion of a privileged relation in terms of an agent's capacity to form, review, revise, suppress and selectively act on a range of her own psychological states. These are states that she may authoritatively claim as her own because she is the one authoring them and authoring them because she authorizes them: She makes them answer to her sense of what is right, or just, or appropriate to think or feel under the circumstances in which she finds herself. Hence, this model also meets the second explanatory charge. An agent who is related in this privileged way to her own psychological states cannot feel alienated from them because they are responsive to her own judgemental determinations. They take the shape they do because of her rational competence, and in that sense may be considered the expression of that competence. It does not seem much of a leap to add that this privileged way of relating to one's own psychological states is the *sine qua non* of well-functioning agency.

In this third section of my paper I want to focus on Moran's specific account of how the agent's deliberations must eventuate in her psychological states if they are to be properly considered *her* states: states that she has rationally authored as against states that are simply inflicted on her, even if she is the one that does the inflicting. In particular, my concern in this section is with the connection Moran makes between having the psychological structure of a rationally autonomous agent in his deliberatively purist sense and meeting an appropriate normative ideal for well-functioning agency. I already indicated in the previous section that I think this condition on authoritative self-knowledge is too stringent. In my view, agents can fail to be rationally or deliberatively autonomous in Moran's sense and still be counted as exerting authorial power over their own psychological states. Moreover, I will shortly argue, *pace* Moran, that such agents are appropriately viewed as psychologically healthy, even admirable, according to our ordinary ways of thinking about these matters. Still, it is hard not to sympathize with the intuition that there is something psychologically more appealing about the agent who meets Moran's more stringent condition on rational autonomy: it seems a better way to be even if not strictly required for first-person authority; hence it may be that we should aspire to this condition. But is this aspiration well-placed?

I think it is not. In this section, I will argue that, far from being the royal road to psychic health, developing deliberative autonomy or 'spontaneity' in Moran's purist sense can be a sign of real psychic disease, indicating a capacity to manipulate oneself through the power of one's own reason into a condition of deep self-deception. Thus, what Moran fails to consider is that there may be warped forms of deliberative spontaneity that are even more corrosive to psychic health than the forms of 'extra-rational' or 'external' manipulation that he indicts. This leads to my conclusion in the following Section 4 that the connection between well-functioning agency and having the sort of psychological capacity required for authoritative self-knowledge is contingent upon a certain degrees of moral-emotional development; and that such development for human beings is practically inconsistent with embracing Moran's rational ideal. This emphasis on moral development represents a reversal in our standard ways of viewing the matter: self-knowledge, so far as it's tied to rational competence and responsible agency, is often presented as an important pre-condition of such development. But if I am right, we cannot achieve the sort of self-knowledge that underwrites a deeper notion of wise and responsible agency until we achieve a certain level of moral development.

I now turn to a consideration of cases mean to challenge these basic elements in Moran's picture: first, that a person's states are *properly* authored by her if and only if they arise spontaneously as the expressive output of her deliberations; and secondly, that the appropriate ideal for well-functioning agency in human beings is the Kantian ideal of rational autonomy. I take for my case studies two well-drawn portraits from George Eliot's novel *Middlemarch*, in which she gives a subtle and extensive portrayal of the ways in which our own powers of agency, for good or ill, are deeply affected by a growing moral wisdom.

Case #1: the Reverend Camden Farebrother, Vicar of St. Botolph's.

The Reverend Farebrother is one of the most appealing characters in Eliot's novel, having achieved some genuine understanding of the varieties of strengths and weaknesses that constitute human character, as well as the temptations and influences to which it is prey. He neither holds himself above others, nor them above himself. His moral understanding is broadly encompassing, making him more ready to forgive than to judge, and if judging, then only with sad acceptance instead of righteous or angry condemnation. On the face of it, he seems rather wise—and just the sort of person one would want for a friend.

Alas, he is not a poster boy for rational autonomy in the Kantian sense. Farebrother is too ready, as Moran might say, to adopt an 'empirical stance' towards himself—viewing himself as a psychological object with a variety of inclinations, impulses, weaknesses and unsatisfied yearnings over which he seems unable to exercise much non-manipulative agential authority. Of course, this tendency is not unreservedly a bad thing, as Moran admits and Eliot makes clear, since it gives Farebrother a degree of moral insight often lacking in others. Hence while he may be too inclined to adopt a third-personal stance towards himself, he also sees others as suffering—though they may not know it themselves—from the same kinds of psychological liabilities that he sees in himself. This is what gives him a particularly forgiving nature. But the costs are there as well, as Eliot shows in the following revealing portrayal:

> The character of the publican and sinner is not always practically incompatible with that of the modern Pharisee [who thanked God that he was not as other men are], for the majority of us scarcely see more

> distinctly the faultiness of our own conduct than the faultiness of our own arguments, or the dulness of our own jokes. But the Vicar of St Botolph's had certainly escaped the slightest tincture of the Pharisee, and by dint of admitting to himself that he was too much as other men were, he had become remarkably unlike them in this—that he could excuse others for thinking slightly of him, and could judge impartially of their conduct even when it told against him.
>
> 'The world has been too strong for **me**, I know,' he said one day to Lydgate. 'But then I am not a mighty man—I shall never be a man of renown. The choice of Hercules [of duty before pleasure] is a pretty fable; but Prodicus makes it easy work for the hero, as if the first resolves were enough .... I suppose one good resolve might keep a man right if everybody else's resolve helped him.'
>
> The Vicar's talk was not always inspiriting: he had escaped being a Pharisee, but he had not escaped that low estimate of possibilities which we rather hastily arrive at as an inference from our own failure. Lydgate thought that there was a pitiable infirmity of will in Mr Farebrother. (Eliot 1996: II.xviii 174–5)

Farebrother's 'infirmity or will' or tendency to lapse into taking an empirical stance towards his own character is certainly something of a liability when it comes to exercising his own agential powers. But in the end, it is not clear just how much of a liability this is. For it is not that Farebrother is completely disabled in making and following through on difficult resolutions, despite his self-characterization. Indeed, he makes a positively heroic choice of duty before pleasure—or, more accurately, of self-sacrifice before the sacrifice of others—when he pleads the romantic cause of Fred Vincy, a young friend who entrusts him with this task, to Mary Garth, the woman he secretly loves and had hoped to marry himself. Fred is a charming, though very weak-willed young man who needs all the help he can get so as not to lose sight of his better nature. In particular, Fred is devotedly and dependently in love with Mary and has been doggedly loyal to her for many years. However, Mary refuses to have anything to do with Fred until he finds the strength of character to do something useful with his life. A compromise is reached with Farebrother's pleading: Mary does plight her troth to Fred but only on condition that he stick by his own resolution to give up his feckless ways. He takes some steps in this direction, but is soon tempted astray by the pleasures of gambling. Hearing of this, Farebrother is faced once again with the question of whether to come to Fred's aid by reminding him that he is likely to lose Mary through his activities, thus continuing to put Fred's interest ahead of his own. His noble impulses do win out in the end, but only after a great deal of internal struggle. More to the point, they win out because Farebrother understands the nature of his own counter-inclinations with regard to keeping the resolution to support Fred and Mary active in himself. Hence, he falls back on strategies of self-regulation and control, the most impressive of which is to enlist Fred's own support for keeping him on track by making Fred aware of the struggle within himself between his loyalty and friendship to Fred and his persisting desire to win Mary for himself by letting Fred go to the dogs: 'I had once meant better than that, and I am come back to my old intention. I thought that I could hardly *secure myself* in it better, Fred, than by telling you just what had gone on in me. And now do you understand me? I want you to make the happiness of her life and your own, and if there is any chance that a word of warning from me may turn aside any risk to the contrary—well, I have uttered it' (Eliot 1996: VII.lvvi 635–6).

The heroic sacrifice is made—or, rather, as Farebrother tells himself he has managed 'a very good imitation of heroism'. But must we concur in his own modest judgement of this act simply because he needed to fall back on strategies of self-regulation to prop up his resolution? That seems a most unfair assessment of his reasoned agential contribution to keeping his psychological states in the condition that he means them to be.[9] Moreover, it seems equally unjust to make the judgement that he is somehow psychologically dysfunctional because of this need, especially in light of the deep admiration we are wont to feel for his action and not just under the general description of putting Fred's interest ahead of his own. It is his method, I think, as well as his intent that excites our admiration. For in so exposing himself to Fred, Farebrother has simultaneously done quite a bit in Fred's service. First, he has warned Fred of an impending danger by alerting him to the existence of a potential rival for Mary's affections—a rival whom she would have very good reason to prefer over a self-preoccupied, irresponsible young man if that young man should fail to change his ways. Secondly, by informing Fred of his own interest, he has refused to remain paternalistically complicit in Fred's tendency to slough off his own agential responsibilities by thoughtlessly relying on others to make things come out right for him. Thirdly, by making clear the nature his own internal struggle, Farebrother has modelled for Fred what it is be an agent who experiences temptations and self-conflict, but who nevertheless finds ways around his own weaknesses with respect to the resolutions he has made—in this case, by exposing himself to, and so enlisting the support of, others. (As Farebrother himself insightfully remarks: 'I suppose one good resolve might keep a man right if everybody else's resolve helped him'.) Finally, by making Fred come to a full understanding of the situation, he has pricked Fred's admiration and sympathy for someone other than himself, thereby undermining Fred's egoistic tendency to think of the world and everything in it in terms of his own needs and interests. In this way, he inspires in Fred the advent of genuinely other-regarding concerns, giving Fred a new way to strengthen his own resolution and develop his own character. As Eliot writes:

> Fred was moved quite newly. Some one highly susceptible to the contemplation of a fine act has said, that it produces a sort of regenerating shudder through the frame, and makes one feel ready to begin a new life. A good degree of that effect was just then present in Fred Vincy . . . .

> 'I shall never forget what you have done,' Fred answered. 'I can't say anything that seems worth saying—only I will try that your goodness shall not be thrown away'. (Eliot 1996: VII.lxvi 636)

Given the multiple goods that arise from this interaction, the fact that Farebrother practices such strategies of self-regulation in light of what he knows about himself seems nothing less than a form of psychological and moral excellence. Still, is there not a better way of being, psychologically speaking? As I said earlier, I am drawn to the intuition that, despite this exemplary moment, there is some overall weakness in Farebrother's character, an incipient passivity that perhaps is best characterized along the lines that Moran proposes. On this account, Farebrother's tendency to adopt an empirical stance towards his own character would amount to a fundamental evasion of the responsibilities of agency, specifically the responsibility for authoring his own mind in accord with his own reason. It is a clever sort of evasion since it works under the guise of being 'psychological realistic' about himself and, therefore, responsible in a certain sort of way (Moran 2001: 81). Nevertheless, as Moran argues, adopting the empirical stance too often and too readily serves as a distraction: it leads an agent to focus less on the actual reasons for believing, desiring or intending something and more on one's own liabilities to fall away from the dictates of reason. This can become a vicious cycle. For the more one distracts oneself from the reasons for one's beliefs, the less powerful one's deliberative determinations become since it is the deliberative focus on one's reasons for believing that empower those determinations in the first place. And, of course, the less powerful one's determinations become, the more reason one has to worry about one's own capacity to think and act in accord with them. Hence, the agent who has a persistent tendency to adopt the empirical stance persistently undermines his own deliberative powers, hence his own capacity to be a rationally autonomous being. Did Farebrother not have such a tendency, he might have accomplished more in his life and required less by way of his own effortful manipulations to govern his own mind. That, at least, is the intuition that favours a character more deliberatively well-ordered than Farebrother appears to be. But is this intuition fully credible?

Case #2: Nicholas Bulstrode, Evangelical Middlemarch Banker.

Turning to *Middlemarch* again, we find an array of characters that suffer no 'infirmity of will' of the sort that plagues Farebrother. They are the visionaries—Tertius Lydgate, Dorothea Brooke, Nicholas Bulstrode—whose passions are directed, as Eliot wryly says, to 'shaping their own deeds and altering the world a little' (Eliot 1996: II.vx 135). In fact, their deliberative energies are wholly consumed by their projects: by their understanding of how the world is and how to make it better. There is much to say about all three of these characters, but Bulstrode in particular stands out as a kind of mirror image of Farebrother.

To begin with he has no doubts at all about his own powers of agency and their proper application. Having discovered early on his evangelical potential, he soon became convinced that he was designated by God for 'special instrumentality'. And though we're introduced to him rather late in his career when he is already a rich and powerful banker in the town of Middlemarch, his religious zeal has suffered no diminution. On the contrary, his 'serviceableness to God's cause' has been the lifelong standard to which he has appealed in all his deliberations, using this as the measure of right reason and sound conduct. Indeed, so devoted is he to this one standard and so accustomed is he to its appeal, that the conclusions he reaches in deliberation are as spontaneous and psychologically effective as anyone might wish who aspires to a condition of rational autonomy.

And, yet, the trajectory of Bulstrode's life has been a disturbing one, as Eliot so poignantly reveals to us. Eager for missionary work in his unblemished youth, she tracks his career as he transformed by degrees into a fence for stolen goods, a liar and a cheat, a petty tyrant in his business and family dealings, and finally something close to a murderer when he turns a blind eye to a servant's ignorant mismanagement of an ailing enemy's care for which he is responsible. Yet all of this is done with an impenetrable self-righteousness that would be hard to credit were it not so psychologically realistic. Here is Eliot's compelling portrait of the workings of Bulstrode's mind:

> The service he could do to the cause of religion had been through life the ground he alleged to himself for his choice of action: it had been the motive which he had poured out in his prayers. Who would use money and position better than he meant to use them? Who could surpass him in self-abhorrence and exaltation of God's cause? And to Mr. Bulstrode God's cause was something distinct from his own rectitude of conduct: it enforced a discrimination of God's enemies, who were to be used merely as instruments, and whom it would be well if possible to keep out of money and consequent influence. Also, profitable investments in trades where the power of the prince of this world showed its more active devices, became sanctified by a right application of the profits in the hands of God's servant. (Eliot 1996: VI.lxi 582)

Reflecting on his character, Eliot comments:

> There may be coarse hypocrites, who consciously effect beliefs and emotions for the sake of gulling the world, but Bulstrode was not one of them. He was simply a man whose desires had been stronger than his theoretic beliefs, and who had gradually explained the gratification of his desires into satisfactory agreement with those beliefs. If this be hypocrisy, it shows itself occasionally in us all. (Eliot 1996: VI.lxi 581)

Of course, Bulstrode does excel at what we might rather call *rationalization*. That is to say, he is particularly adept at putting his reason to work in the service of a feature of his psychology to which he is determinedly blind—viz., 'his immense need of being something important and predominating' (Eliot 1996: VI:lxi 582). Indeed, it is because he is so blind to this need that it is invisibly transformed into a perceptible feature of Bulstrode's world—not as need, of course, but as the very condition that would satisfy this need: he becomes God's chosen instrument. With this 'perception' firmly in place, Bulstrode has no difficulty at all in responding to the dictates of his reason. For now his reason is geared to authorize in him—and for him alone—whatever ambitions and temptations he experiences since these must be connected with God's design. Hence, the appearance of hypocrisy: For he can readily condemn in others the selfsame attitudes and actions that he authorizes in himself. In them they are evil and contemptible, whereas in him there is a divine purpose that they ultimately serve.

Now I take it no one would argue that, because of the spontaneity with which Bulstrode's psychological states respond to the dictates of his reason, he is a model of well-functioning agency. Something has gone deeply wrong in him; so wrong, in fact, that he has become a victim of the authority with which reason speaks in him as it leads him into greater depths of self-deception and moral inequity. It is tempting to say that he suffers from a warped form of rational autonomy: indeed, his rational deliberations are made all the more powerful and effective because of the way they have been hijacked by psychological forces that are completely invisible to him. But if this is so, and if Bulstrode's case is exemplary of a more general cognitive liability found in all of us, then Moran's Kantian ideal of 'handing over the question of one's beliefs or intentional action to the authority of reason' cannot be an entirely happy one. At the very least, we must give sober consideration to how vulnerable we are to such corruptions of reason and how best they can be guarded against. As Eliot wisely observes of Bulstrode's cognitive condition:

> This implicit reasoning is essentially no more peculiar to evangelical belief than the use of wide phrases for narrow motives is peculiar to Englishmen. There is no general doctrine which is not capable of eating out our morality if unchecked by the deep-seated habit of direct fellow-feeling with individual fellow-men. (Eliot 1996: VI:lxi 582)

4. The Moral Development of First-Person Authority

In the previous section, we focussed on Moran's account of the relation a person must bear to her own psychological states if she is to manifest the kind of first-person authority that is characteristic of well-functioning agency. According to Moran, such an agent must be psychologically well-ordered in the sense that her psychological states arise spontaneously from her deliberations. For this means that she need not resort to any self-regulative or self-manipulative strategies directed at herself—now from the stance of considering herself a limited empirical being—in order to bring her psychological states into line with the dictates of her reason. However, this condition does not seem to be necessary for our judging a person to have the sort of psychological capability that makes for well-functioning agency, as Farebrother shows. Nor does it seem to be sufficient, since someone like Bulstrode may be deeply self-deceived about the true nature of his agential motivations and activities. In this final section, I want to push this criticism a step further by discussing the moral implications of the two cases we have considered. For it seems clear that Farebrother's tendency to adopt some kind of empirical stance towards himself is not unrelated to the moral admirability of his character. Likewise, Bulstrode's inability—perhaps even refusal—to disengage from what Moran calls the transcendental perspective of agency is not unrelated to the moral disreputability of his character. But how exactly are we to understand these connections? And what implications will this have for fashioning a morally responsible conception of first-person authority? In order to answer these questions, it will help to consider the moral-psychological structure of these characters, as Eliot does, from a developmental perspective.

Reflecting on our common human situation, Eliot observes that 'we are all of us born in moral stupidity, taking the world as an udder to feed our supreme selves' (Eliot 1996: II.xxi 198). Such stupidity is not, of course, malicious. It is simply the result of a kind of native egoism that invariably distorts our perceptions of the world around us. Dramatising this condition in a wonderful passage, Eliot writes:

> An eminent philosopher among my friends, who can dignify even your ugly furniture by lifting it into the serene light of science, has shown me this pregnant little fact. Your pier-glass or extensive surface of polished steel made to be rubbed by a housemaid, will be minutely and multitudinously scratched in all directions; but place now against it a lighted candle as a centre of illumination, and lo! the scratches will seem to arrange themselves in a fine series of concentric circles around that little sun. It is demonstrable that the scratches are going everywhere impartially, and it is only your candle which produces the flattering illusion of a concentric arrangement, its light falling with an exclusive optical selection. These things are a parable. The scratches are events, and the candle is the egoism of any person now absent . . . '. (Eliot 1996: III:xxvii 248)

Lifting ourselves out of this native condition of moral stupidity is thus for Eliot our common developmental challenge: 'to conceive with that distinctiveness which is no longer reflection but feeling . . . [that others have] centre[s] of self, whence the lights and shadows must always fall with a certain difference' (Eliot 1996: II:xxi 198).

Exploring Eliot's views a little further, we might ask a number of questions relevant to our topic of concern: How can we escape from this native condition of 'moral stupidity', and why does it matter for our becoming well-functioning authoritative agents? Why is *feeling*, as opposed to mere (intellectual) reflection, an important component in coming to see others as having 'centres of self? How is coming to see others in this way pertinent to the development of a morally relevant *self*-understanding? And, finally, what impact will this development have on our capacities to act as responsible, authoritative agents in our interactions with others and with the world? Here are the answers I think Eliot would give.

To escape from our native condition of moral stupidity, we must redirect some of the affective energy that we are naturally inclined to pour into sustaining ourselves as the 'proper and true' centre of things. This is critical for developing into psychologically capable agents, in Eliot's view, because it is this energy—and the distorting cognitive and perceptual frame it induces—that has the power to hijack our reason, creating self-serving biases in our thinking even as we imagine ourselves to be reasoning objectively about the world and our situation in it. We saw this process at work in Bulstrode, whose agential powers are dramatically supported and enhanced by his affectively driven, ego-centred conviction that he operates under an objective perception of the true world order. Of course, Bulstrode is not the only character who suffers from this affective-perceptual liability; nor is his manner of suffering it the only possibility. For instance, we see the very same biasing process at work in weak-willed Fred Vincy, mentioned passingly in the previous section.

Now this comparison between Fred and Bulstrode may seem surprising. After all, Fred appears to be the very antithesis of Bulstrode in so far as his agential powers are continually *undermined* by his 'infirmity of will'. Again, as the very antithesis of Bulstrode, Fred seems chronically unable to adopt the transcendental perspective of agency, viewing himself as a person who can get nowhere on his own initiative. (Hence, his constant dependence on Mary's reinforcing love: 'I will never be good for anything, Mary, if you will not say that you love me' (Eliot 1996: II:xiv 130).) So why not liken Fred to Farebrother, given that both characters are inclined to distance themselves from their own psychological states by adopting an 'objective' attitude towards themselves?

This would miss the deeper lesson that Eliot is trying to drive home. As the novel makes clear, Fred's 'tendency' is quite unlike Farebrother's and rather more like Bulstrode's in so far as it is compulsively ego-driven—i.e. both Fred and Bulstrode live in worlds that are egocentrically structured around their own needs. Of course, in Fred's case, his overwhelming need is to offload the responsibility for his life onto others. Hence, Moran, who gives a brief but telling analysis of Fred's character in his own work, is quite right to insist that adopting an empirical stance towards oneself *can* function as a way of evading one's agential responsibilities (Moran 2001: 187–94). But, as we now see, serving this function is not a feature of the stance itself but is rather a feature of the egoism that drives agents to use whatever means are necessary to satisfy the needs most dominant in them. In Bulstrode's case, his egocentric needs are better served by an opposite, though equally compulsive adherence to the transcendental perspective of agency.[10] Thus, to the extent that our capacity to function well as agents depends on having a psychological capacity for undistorted authoritative self-direction, it seems a first and primary concern should not be with the sort of stance we adopt towards our own psychological states, but rather with the *way* in which we adopt that stance. And this in turn will depend on the stage of moral-emotional development we have reached. In particular, it will depend on our overcoming an entirely natural, though inevitably corrupting egocentric orientation to the world.

So how is this native egoism to be developmentally overcome? In Eliot's view, because such egoism is affectively sustained, our best bet for overcoming it is to redirect some of the affective energy we naturally focus on ourselves onto the plight of others. The consequences of this reorientation will go deeper than a mere intellectual acknowledgement that others' physical or even psychological situation differs from ours, since it will allow us to *experience* the world as it is experienced by them; it will allow us to see the world as it comes to them, distorted through the lens of their affectively-laden perceptions.[11] One immediate consequence of this sympathetic engagement with others is thus coming to see them more nearly as they are in themselves—namely, as 'centres of narrative gravity' (Dennett 1991: Chapter 13) with a powerful, albeit for them largely invisible, capacity to shape and distort the perceptual and cognitive space in which they operate.

This dawning understanding of others signals a critical turning point in our own moral and epistemic development. For now we are in a position to see ourselves in an entirely new light—namely, as one other such being whose world—the world of *our own* experience—is projectively distorted by the fluctuating hopes and fears, desires and needs that constitute our empirical selves (our facticity, to use Moran's term). In short, we come to see ourselves as one sun among many, always in danger of thinking and acting under the terrible power that reason becomes if trapped in a universe of our ego's own distorted making. Thus, from a developmental perspective, it seems our best protection—indeed, our *only* protection—against an ego-driven corruption of reason is to cultivate an *allocentric* capacity to see ourselves as we see others—namely, as empirical subjects whose psychological states are responding to a variety of influences that are largely invisible from a naïvely egocentric first-person point of view. And this in turn suggests that Farebrother's susceptibility to adopting an (allocentric) empirical attitude towards himself is not mere weakness of will or desire for evasion but represents, more significantly, a substantial developmental achievement in moral and epistemological terms.

But there are also agential costs to this achievement, as both Eliot and Moran make clear. We are thus left with the following developmental dilemma: On the one hand, we can agree with Moran that adopting the transcendental perspective of agency is essential for individuals to manifest the sort of first-person authority that is characteristic of well-formed agency. For this is the perspective from which agents understand themselves to have psychological states that are open to their determination; it is the perspective from which they assert their freedom—and their responsibility—to think and act with first-person authority. On the other hand, we can agree with Eliot that morally developed agents are precisely those agents who have acquired the capacity to see themselves from an appropriately allocentric empirical perspective. These are the agents who are able to protect themselves from the most egregious forms of egocentrically induced self-deception, but only—as Farebrother's case makes clear—by cultivating self-critical habits of mind aimed at undermining any easy well-insulated adoption of the transcendental perspective of agency. Yet, problematically, these are also the agents whose biggest risk is a loss of authorial nerve: a conviction that their best and most moral choice is to undermine their own first-person authority with the sort of on-going irresolution characteristic of a merely spectatorial, objectifying knowledge of their own empirical selves. As Eliot so poignantly remarks: 'If we had a keen vision and feeling of all ordinary human life, it would be like hearing the grass grow and the squirrel's heart beat, and we should die of that roar which lies on the other side of silence. As it is, the quickest of us walk about well wadded with stupidity' (Eliot 1996: II:xx 182). Perhaps the quickest, but not the best—if Bulstrode's character is any indication. So how is this developmental dilemma to be resolved?

In my view the solution is at hand, though it will not sound very appealing to those who are theoretically committed to a purist ideal of deliberative autonomy. For that ideal is precisely what has been called into question—not *tout court*, it must be emphasized, but only for creatures like us, with our particular, natively given cognitive liabilities. Of course, it is pleasing to think that we could overcome these liabilities once and for all; hence, that a deepening moral understanding would go hand in hand with becoming the kind of character that is fully virtuous in something like an Aristotelian sense—viz., a person who thinks and acts as she should without having to struggle against deliberatively recalcitrant needs, desires and temptations. But given the wide range of human frailties to which all of us are prey, and given—beyond that—our native susceptibility to overlooking such frailties in our own case,[12] it is far more likely that persistent deliberative success in silencing recalcitrant needs, desires and temptations will signal the sort of rationalizing 'quickness' that is the hallmark of egocentrically induced self-deception.

To guard against this possibility, the morally wise agent is continually ready to step back from her own character, disempowering the authorizing voice of her own reason to some extent in order to make possible a more objective assessment of her own appetites, needs, weaknesses, and reactive impulses as psychic forces potentially shaping, as much as being shaped by, her own deliberative processes. Of course, as Moran makes clear, the morally wise agent cannot rest content with the quietude of such empirical sophistication, since this would amount to denying the responsibilities—and the privilege—of her first-person authority. Hence, such an agent must re-empower her own reason, but in a more complex way. She must think and act resolutely, deliberating not only about what *is* to be believed, desired or intended, but also about how best to keep the potentially derailing forces of her own psyche in line, given that she now understands how these are likely to emerge under conditions of stress, temptation and other sorts of cognitive-emotional pressures. In effect, she must trade in the purist ideal of rational autonomy for a different sort of ideal—we might call it an *agonistic* ideal—whereby she honours the fact that well-functioning first-person authority on a human scale should expectedly involve a certain amount of extra-deliberative struggle on the agent's part to govern, regulate or control wayward tendencies that do not fit with the psychological states she deliberatively stands behind or endorses.

My conclusion, succinctly stated, is as follows: A morally mature agent, a morally wise agent, is one who understands the peculiar responsibility she has for making and maintaining her own psychological states. But such an agent is also one who understands the frailties of human nature, including those rationalizing tendencies that subvert reason itself, and so abjures the ideal of rational autonomy in favour of the more human-sized ideal of rational agonism. I think this position strikes the right balance between being either overly optimistic or overly pessimistic about our capacities as morally responsible authoritative agents. But I imagine there will be critics on either side. So I close with a few final reflections aimed at addressing each of these constituencies.

For those who find this conclusion too pessimistic, the complaint may be that if we do not aim high enough as deliberative agents, we will achieve less in this dimension than we otherwise could. The point of endorsing an (unrealisable) regulative ideal is precisely to encourage individuals to develop their current capacities so far as they are able. Therefore, it seems downright mistaken to replace a genuine ideal, such as rational autonomy, with a downgraded alternative, such as rational agonism, that simply accepts limitations right from the start.

My answer to this concern is that the ideal of agonism is no less of an ideal for being agonistic: it has built into it the (unrealisable) ambition of deliberatively determining the shape of our own psychological states; it is simply more modest about the variety of means we ought rationally to employ in order to pursue that ambition. Moreover, built into this ideal is the recognition that becoming psychologically realistic about the deliberatively recalcitrant forces operating in ourselves is itself a demanding (and unrealisable) ambition and, further, that taking appropriate steps to counteract these forces, without indulging their power too much, can be strategically, affectively and morally challenging: witness both the difficulty and the excellence of Farebrother's solution to his own 'infirmity of will'. So it is simply not the case that the agonistic ideal fails to be a very demanding form of regulative ideal. Still, it is what I have called a human-sized ideal. And by this I mean to call attention to the fact that some ideals can look very good on paper, but translate rather badly into reality. I propose that the ideal of rational autonomy is one such ideal. In an unqualified sense, it may represent the best way an agent could be. And by this token, we may accept it as an appropriate *constitutive* ideal even for human beings. But the 'best' can sometimes be an enemy of the 'good', and the path of wisdom is to understand when such a situation is likely to obtain. What I have tried to show in this paper is that the ideal of rational autonomy is a poor sort of *regulative* ideal for human beings, not because it is practically unobtainable, but because it encourages the kind of deliberative spontaneity that, for us at any rate, can so often mask tendencies towards rationalising self-deception.[13]

But what about the opposite sort of concern, that my conclusion is too optimistic? The worry here is that I have too readily equated a certain moral maturity with embracing the agonistic ideal for first-person authority. Surely there may be individuals who work to realise this ideal, and yet act badly from a moral point of view. Indeed, they may act badly in part *because* they embrace the ideal. They have morally sound feelings or concerns that they cannot deliberatively silence. Yet they come to regard these internal misgivings, not as morally sound, but rather as morally objectionable, indicating forms of psychological weakness that they simply need to work around. (Think here of some of the American soldiers who, against their own feelings of anxiety or revulsion, ended up torturing prisoners in Iraq's Abu Ghraib.)[14]

I grant this is a real worry and I have no recipe for overcoming it. However, the problem is not so much that these individuals cannot deliberatively silence misgivings or second thoughts, but rather that they end up deliberatively endorsing morally iniquitous behaviour, that this finally constitutes their *best judgement* about what is to be thought or done. In short, they fail as moral agents, not at the level of regulating their own psychological states, but rather at the level of deliberation itself. So I think this sort of case does not defeat the main thesis. Once again my claim is simply this: that embracing an agonistic ideal is the appropriate outcome of a certain degree of moral-psychological development. Of course, it does not follow from this that embracing such an ideal signifies that such development has taken place; nor does it follow that embracing such an ideal is sufficient on its own for moral wisdom. There are many ways to fail at being a good moral agent, and merely avoiding certain kinds of pitfalls in one's own deliberative processes does not entail avoiding them all. Still, this objection raises an important point—viz., that when a person is divided against herself, when she experiences persistent feelings of reluctance or anxiety about what her reason endorses, these feelings may be telling her something that she needs to pay attention to at the level of deliberation (Jones, K. 2004). That said, there is no algorithm for endorsing or rejecting any anxiously delivered second thoughts; and the morally wise agent is one who understands that her own psychological resistances cannot be expected to wear their good or bad credentials on their sleeves.[15]

Victoria McGeer
Department of Philosophy
Princeton University
Princeton, NJ 08544 1006
USA
vmcgeer@princeton.edu

NOTES

[1] My thanks to Carrie Jenkins for putting this concern to me so succinctly.

[2] Here is a brief characterization of the difficulties encountered by the epistemic approach, even on its own terms. It is generally acknowledged that privileged knowledge

of our own mental states is limited to those states that are apt for folk-psychological characterization. These are the states we normally attribute to others in the course of our day-to-day interactions and normally claim for ourselves. They are the states that enter into our discourse of normative assessment regarding our own and other's behaviour. They are the states in accord with which we dole out praise and blame, feel resentment and other reactive attitudes. In short, they are the intentional states that underpin our notions of agency and responsibility: beliefs, desires, intents, emotions and so on. Naturally, this leaves out a whole range of cognitive or neuro-computational states and processes that operate below the level of consciousness—indeed, below the level of what ever could become conscious. And this may seem to be no problem. But, in fact, it is a problem for the epistemological approach, since social psychologists have persuasively shown that, even under the most quotidian circumstances, these states and processes can exert considerable influence on our behaviour, contrary to what we ourselves think about the matter. See, for instance, studies conducted by Nisbett and Ross 1980, Nisbett and Wilson 1977. What these studies show is that first-person phenomenology of 'privileged access' is no guarantee of epistemic reliability; worse, they strongly suggest that our epistemic position vis-à-vis our own psychological states is not reliable enough to warrant a default presumption of first-person authority. Hence, if epistemic superiority is put forward as the fundamental ground upon which the presumption of such authority is based, it becomes hard to resist the arguments of philosophers and psychologists who claim that first-person authority is nothing more than a subjectively grounded and culturally reinforced illusion. See, for instance, Churchland 1979 and Gopnik 1993.

[3] Akeel Bilgrami, who has independently proposed a very similar thought-experiment, argues that an entirely passive subject would actually be worse off than this discussion implies. For, on Bilgrami's view, there is no reason to count such a subject as having any (first-order) thoughts at all. To have thoughts is to have contentful intentional states that do not just come and go without rational explanation, as this scenario seems to imply. Rather, they come and go because the thinker is actively weighing evidence, making inferences, checking for consistency and so on. In other words, having thoughts implies thinking, which in turn implies a subject who is active with respect to the formation of her first-order intentional states. And this in turn implies a subject who has (authoritative) knowledge of her own intentional states. See Bilgrami 1998: 234–41. This insistence on a conceptual linkage between having thoughts at all and having authoritative self-knowledge to fits very well with the agency model of self-knowledge discussed in Section 2.

[4] Other important points of reference include works by Bilgrami 1998, Shoemaker 1996, Burge 1996, Dennett 1987, Dennett 1991, Wittgenstein 1958 and Sartre 1963.

[5] Cf. Moran 2001: 59, 'There is a . . . dynamic or self-transforming aspect to person's reflections on his own state, and this is a function of the fact that the person himself plays a role in formulating how he thinks and feels'. Akeel Bilgrami endorses a similar point, as discussed in note 3.

[6] It may also sound objectionably rationalistic, as if the states we come to endorse are always the result of explicit deliberation. Again, as Moran 2001: 116 insists, the impression is misleading: 'For a desire to belong to the "judgement-sensitive" category it is, of course, not necessary that it be *formed* as the result of deliberation. For very few of our desires come into existence as the conclusion an explicit exercise of practical reasoning. Equally, however, very few of our beliefs about the world arrive as the conclusion of any explicit *theoretical* reasoning that we undertake. It is nonetheless essential to the category of belief that a belief is a *possible* conclusion of some theoretical reasoning . . . Similarly, what is essential for a desire to count as "motivated" in the relevant sense is for it to be the

possible conclusion of some practical reasoning'. I merely add that such beliefs and desires are under our *virtual* rational control: when we come to express what we believe and desire, spontaneously as it may be, we willy-nilly call the contents of these states to our own explicit attention, and in doing so create the opportunity for such claims to be deliberatively reviewed in light of our other commitments. For further discussion, see also Pettit 1996 and McGeer and Pettit 2002.

[7] In this context, it is interesting to compare the distinction Akeel Bilgrami makes between what he calls states that are 'fully intentional' (such as beliefs) and mere 'dispositions' in Bilgrami 1998: 240–1. Intentional states, in Bilgrami's view, are more like commitments: e.g., a person is appropriately attributed the belief that p just in case she is committed to the truth of p and shows that commitment by accepting criticism for not thinking or acting in ways that are commensurate with being so committed. Dispositions, by contrast, are those states that constitute a person's overall behaviour-inducing psychological profile. See also Levi 1980. Given Bilgrami's characterization, his 'intentional states' (beliefs and desires) are more like my 'judgements', and his 'dispositions' are more like what I call 'intentional states'. This may be more of a terminological difference than a substantive one. But, as argued in McGeer and Pettit 2002, I think there are good reasons to continue to use the fully intentional language of 'belief' and 'desire' to characterize our underlying dispositions, while still agreeing with Bilgrami that, as humanly instantiated, such dispositions are amenable to normative regulation. For a compatible view of beliefs as normatively regulated dispositions, see also Schwitzgebel 2005.

[8] Cf. Moran 2001: 117, ' . . . a person thinking that it's getting late or his hoping for rain are not *effects* he produces, even when they result from a process of thinking on his part. Instead, such attitudes are constituents of his thinking and are thus more analogous to the act of pinching than to the sensation produced by that act'.

[9] For discussion of a similar sort of case, see Pettit and Smith 1993 on the internal underpinning of reason.

[10] At various places in his book—see, for instance, Moran 2001: 160–4—Moran suggests that the ability to adopt both the transcendental and the empirical perspective on oneself is essential for one's psychological well-being since each perspective delivers truths about the self that are inaccessible from the other perspective. Thus, to avoid or ignore either one of these perspectives is to engage in a distinctive kind of moral evasion. In abstract terms, I completely agree with this claim as my discussion here is meant to show. But I disagree with the way Moran characterizes the distinctive form of moral evasion that comes with avoiding or refusing to adopt the empirical stance of psychological facticity—a problem that I have associated with Bulstrode's overweening psychological self-mastery. In Moran's view, by contrast, this kind of evasion involves refusing to face or acknowledge the 'empirical reasons' for which one has 'lost the right' to expect one's deliberative powers to be effective in determining one's psychological condition, thus making one's assumption of agency a 'mere sham' based on nothing but 'wishful thinking' about the power of one's own resolutions (Moran 2001: 163; cf. 81). I agree with Moran that this constitutes a distinctive kind of pathology, but I think it is better represented as one more manifestation of the wishful desire to locate the (real) burdens of agency elsewhere. Thus, an agent may make some fanciful pretence of engaging the transcendental perspective of agency, only to justify a self-exculpating retreat to the empirical stance of psychological facticity when she 'discovers' her resolution to fail. Indeed, this is just the kind of seesawing one sees in the character of Fred Vincy. For further discussion of Fred's 'wishful' character particularly in comparison with Bulstrode's, see McGeer 2004.

[11] I imagine Moran would agree with Eliot that a mere intellectual acknowledgement that others have centers of self may not go deep enough to undermine the distorting form of egoism that is at issue here. Such egoism is akin to an emotional attitude; and, as Moran 2001: 181 says, distinguishing these from mere beliefs, 'an emotional attitude constitutes something closer to a total orientation of the self, the inhabiting of a particular perspective'.

[12] In this connection, it is interesting to note how deeply rooted social psychologists take the 'correspondence bias' (or fundamental attribution error) and a mirroring 'actor-observer bias' to be. According to these biases, agents are far more likely to explain others' actions in terms of standing traits or dispositions, rather than in terms of situational factors, and their own actions, by contrast, in terms of situational factors, rather than standing traits or dispositions. For discussion, see Jones and Harris 1967, Jones and Nisbett 1971 and Ross 1977. While the reason for these biases is much debated, my hypothesis is that they reflect, on the positive side, our first-personal capacity to maintain the self-determining perspective of agency, but also on the negative side, a relative inability to see the standing traits or dispositions that powerfully affect our own behaviour.

[13] My thanks to Karen Jones for suggesting that I distinguish between regulative and constitutive ideals in order to help clarify my position.

[14] My thanks to an anonymous referee at the *European Journal of Philosophy* for raising this concern.

[15] Earlier drafts of this paper were presented in a number of places, including the annual meetings of the APA (Pacific Division) & the Australasian Philosophical Association, the Ohio State University/Maribor/Rijeka Conference on Regulating Attitudes with Reasons, the Fellows seminar at the University Center for Human Values at Princeton University, and to the philosophy colloquia at various other institutions. I have benefited enormously from my discussions in all of these places, but special thanks go to Michael Bratman, Justin D'Arms, Rachena Kamtekar, Karen Jones, Bojana Mladenovic, Philip Pettit, Laura Schroeter, Michael Smith, and Jan Zwicky.

REFERENCES

Armstrong, D. (1968, 1993), *A Materialist Theory of the Mind*. London: Routledge.

Bilgrami, A. (1998), 'Self-knowledge and Resentment', in C. Wright, B. C. Smith and C. Macdonald (eds), *Knowing Our Own Minds*. Oxford: Oxford University Press.

Burge, T. (1996), 'Our Entitlement to Self-Knowledge', *Proceedings of the Aristotelian Society*, 96: 91–116.

Churchland, P. (1979), *Scientific Realism and The Plasticity of Mind*. Cambridge: Cambridge University Press.

Dennett, D. (1987), *The Intentional Stance*. Cambridge, MA: MIT Press.

—— (1991), *Consciousness Explained*. Boston, MA: Little, Brown & Company.

Eliot, G. (1996), *Middlemarch*. Oxford: Oxford University Press.

Gopnik, A. (1993), 'How We Know Our Minds: The Illusion of First-Person Knowledge Of Intentionality', *Behavioral and Brain Sciences*, 16: 1–14.

Jones, E. E. and Harris, V. A. (1967), 'The Attribution of Attitudes', *Journal of Experimental Social Psychology*, 3: 1–24.

Jones, E. E. and Nisbett, R. (1971), *The Actor and the Observer: Divergent Perceptions and the Causes of Behavior*. New York: General Learning Press.

Jones, K. (2004), 'Emotional Rationality as Practical Rationality', in C. Calhoun (ed.), *Setting the Moral Compass: Essays by Women Philosophers*. New York: Oxford University Press.

Levi, I. (1980), *The Enterprise of Knowledge*. Cambridge, MA: MIT Press.

McGeer, V. (1996), 'Is "Self-knowledge" an Empirical Problem? Renegotiating the Space of Philosophical Explanation', *Journal of Philosophy*, 93: 483–515.

—— (2004), 'The Art of Good Hope', *Annals of the American Academy of Political and Social Science*, 592: 100–27.

McGeer, V. and Pettit, P. (2002), 'The Self-Regulating Mind', *Language and Communication*, 22(3): 281–99.

Moran, R. (1997), 'Self-Knowledge: Discovery, Resolution and Undoing', *European Journal of Philosophy*, 5: 141–61.

—— (2001), *Authority and Estrangement: An Essay on Self-knowledge*. Princeton, NJ: Princeton University Press.

Nisbett, R. and Ross, L. (1980), *Human Inference: Strategies and Shortcomings of Social Judgement*. Englewood Cliffs, NJ: Prentice Hall.

Nisbett, R. and Wilson, T. D. (1977), 'Telling More Than We Can Know: Verbal Reports on Mental Processes', *Psychological Review*, 84(3): 231–59.

Pettit, P. (1996), *The Common Mind: An Essay on Psychology, Society and Politics*, paperback edn. New York: Oxford University Press.

Pettit, P. and Smith, M. (1993), 'Practical Unreason', *Mind*, 102: 53–80.

—— (1996), 'Freedom in Belief and Desire', *Journal of Philosophy*, 93: 429–49.

Ross, L. D. (1977), 'The Intuitive Psychologist and his Shortcomings: Distortions in the Attribution Process', in L. Berkowitz (ed.), *Advances in Experimental Social Psychology.* New York: Random House.

Sartre, J. -P. (1963), *The Problem of Method*. London: Methuen & Co. Ltd.

Schwitzgebel, E. (2005), 'Acting Contrary to our Professed Beliefs', available at http://www.faculty.ucr.edu/~ eschwitz/, Riverside, CA: University of California, Riverside.

Shoemaker, S. (1996), *The First Person Perspective and Other Essays*. Cambridge: Cambridge University Press.

Wittgenstein, L. (1958), *Philosophical Investigations*. Oxford: Blackwell.

Wright, C. (1991), 'Wittgenstein's Later Philosophy of Mind: Sensation, Privacy and Intention', in K. Puhl (ed.), *Meaning Scepticism*. Berlin: de Gruyter.