

## Is There a Mechanism of Critical Reasoning?

Draft as of July 7, 2011

Graham Hubbs

hubbs@uidaho.edu

**NOTE:** *this paper was presented at the University of Utah's Philosophy Colloquium on April 29, 2011. I thank Jim Tabery and everyone at the Department there for the invitation and the opportunity to speak. This paper is presently under revision for journal submission. Comments are welcome, and please do not cite without permission.*

### I. The rationalist account of self-knowledge

There is a trend in contemporary epistemology to understand our entitlement to self-knowledge, in a range of fundamental cases, as resting on our ability to give and to assess reasons. Tyler Burge pursues this approach in his mid-90s papers on self-knowledge, and more recently Richard Moran has followed this approach in his book *Authority and Estrangement*. According to this approach, in the relevant cases the entitlement is uniquely first-personal because it does not rest on second- or third-personal ways of acquiring knowledge, such as testimony, observation, or inference. In this paper, I will refer to this approach to self-knowledge as the *rationalist approach*. The line of thought that supports it is demonstrated by the following example. Suppose that I revise a belief as a result of critical reasoning. Let the belief in question concern whether or not Glenn has told a lie. I start off believing that Glenn hasn't told a lie, but upon review I find that I have decisive reason to believe that he has. Accordingly, I revise my belief about Glenn's honesty. As a result of this change of mind, I not only believe that Glenn has told a lie, but I *know* that I believe that he has told a lie. The latter epistemic state is a bit of self-knowledge. According to the rationalist approach, I am entitled to this bit of self-knowledge by the well-functioning of my ability to give and to assess reasons. I know that I believe that Glenn has told a lie not by passively observing this belief but by actively producing it. This production is uniquely first-personal; *my* belief here results from the activity of *my* ability to reason. Because the production is uniquely first-personal, so too is my entitlement to this knowledge of the resulting belief. A person

may be entitled in this way either by rationally changing a belief, as in the example, or by rationally forming a belief, or by rationally maintaining a belief on the basis of critical review. In all of these cases, a person is entitled in a uniquely first-personal way to her knowledge of her belief due to the well-functioning of her ability to produce and to assess reasons.

To better understand the rationalist approach, let us consider some aspects of Moran's specific version of it in some detail. Moran argues that a person may enjoy a uniquely first-personal entitlement to knowledge of those states and attitudes of hers that can be immediately affected by her assessment of their reasonableness. These states and attitudes include beliefs, intentions, and certain emotions and desires. His line of thought, specifically as it pertains to beliefs, is revealed by the following variation on the example from the introduction. Suppose I am part of a jury whose job it is to determine whether or not Glenn is guilty of lying. If I am a good juror—and let us suppose I am—I will enter the trial with an open mind and allow the evidence to determine whether or not I judge him to be innocent or guilty. Until the trial is complete, my mind is not made up; until the trial is complete, I do not believe that Glenn is innocent, nor do I believe that he is guilty. At some point, though, the trial is complete, and I must make up my mind one way or another. Let us suppose that the evidence decisively points towards his guilt and that I make my judgment accordingly. Based on the evidence, I come to judge, and hence to believe, that Glenn is guilty of lying.

The way the story has just been told, no self-judgment or self-belief is involved; the only object of judgment or belief is Glenn. Let us now alter the story slightly. All of the evidence is in, and I am to decide whether or not Glenn is guilty. This time, however, I make the decision by silently posing the following reflective question to myself: Do I believe that Glenn is guilty? Moran notes that despite what the surface grammar might suggest, to answer this question I do not examine myself for evidence of what I think about Glenn. Instead, I consider exactly the same evidence that I would consider were I to pose the simpler question, Is Glenn guilty? That is, in answering this question concerning what I

believe, I consider exactly the same evidence in this altered scenario as I did in the scenario from the previous paragraph. I thus answer my question affirmatively and silently say to myself, Yes, I believe that Glenn is guilty. Put this way, I have arrived not only at a judgment about Glenn—*viz.*, that he is guilty—but also at a bit of knowledge about myself—*viz.*, that I believe him to be so.

In coming to gain this self-knowledge by thinking not about myself but rather about Glenn, my thinking obeys what Moran calls the “Transparency Condition.” The condition is so-called because for me the question, “Do I believe that Glenn is guilty?” is transparent to the question, “Is Glenn guilty?”. When I gain self-knowledge in a way that obeys the Transparency Condition, the means I have employed towards achieving this self-knowledge are neither observational, nor inferential, nor testimonial. When I come to know what I believe about Glenn, I am not uncovering and observing some belief I already had, for the relevant belief does not exist until I make up my mind about Glenn’s guilt. When I gain this self-knowledge, it is not via an inference I make about myself; again, the fact that would be known by drawing such an inference does not exist until I make up my mind about Glenn. When I gain this self-knowledge, it is not by one part of me listening to another part tell the first part what the second part already believes. We can gain knowledge of the world beyond us, including the minds of others, through observation, inference, and testimony. When we gain self-knowledge through conformity with the Transparency Condition, none of these means is employed, so the entitlement is neither one of observation, nor inference, nor testimony.

This last example shows, I hope, why I want to label Moran’s account one that follows a rationalist approach to self-knowledge. It understands the entitlement I have to my knowledge of my belief about Glenn as based on my ability to produce this belief immediately through my assessment of what is most reasonable. Only I can produce a belief of mine immediately through my assessment of what is most reasonable, so my knowledge of the belief is entitled in a uniquely first-personal way. Moran readily acknowledges that the vast majority of our beliefs are not formed in the consciously

explicit manner with which I form my belief about Glenn. Nonetheless, he argues, a person enjoys this uniquely first-personal sort of entitlement whenever she is in a *position* to maintain or to alter her belief as an immediate result of exercising her ability to reason. Moran describes this position as “the stance of agency.” This stance is perhaps most clearly defined by the following conditional counterfactual: a person is in the stance of agency towards some belief of hers if, were she to be presented with adequate reason for revising the belief or giving it up, she would on that basis immediately revise it or give it up. When a person occupies this stance towards a belief, her thought obeys the Transparency Condition. If, while in the stance of agency, she discovers some belief of hers to be unwarranted, she will, in the normal case, immediately revise the belief. When a person lets reason determine her belief in this way, her entitlement to knowledge of the belief is of the uniquely first-personal sort described above.

As already noted, Moran argues that a person can occupy this stance of agency not just towards beliefs but towards any state or attitude that can be immediately rationally affected by reasoning. Take, for example, my desire and intention to go to Memphis to visit the Stax Museum of American Soul Music. Suppose the only reason I want to go to Memphis is to visit the museum, and suppose the museum closes. Before it closes, both my desire and intention to visit Memphis are made reasonable by the fact that I will be able to visit the museum if I go there. If I come to learn that the museum has closed, however, it is rational for me immediately to alter both my desire and intention to go to Memphis. My reason for going there no longer exists, so it is reasonable that my desire and intention to go there no longer exist either. If I alter my desire and intention to go to Memphis immediately because I no longer have a reason to go there, I have rationally changed my mind in a way that parallels the way in which, in the initial example from the introduction, I change my mind about Glenn. This being so, my knowledge of my desire and intention to go (or, if I change my mind, not to go) to Memphis is entitled in the same way as is my knowledge of my belief about Glenn. My knowledge of my desire and intention is entitled not by observation, inference, or testimony, but by the well-

functioning of my ability to reason.

Moran acknowledges that this rational, uniquely first-personal sort of entitlement is not the only sort that a person can have to a bit of self-knowledge. Sometimes, our entitlement to self-knowledge is of the same sort we can have to knowledge of the minds of others. Moran describes the epistemic position that gives rise to this sort of entitlement in a variety of ways: he uses the heads “the stance of appraisal” (p. 192), “the empirical perspective” (see pp. 77-83), “the stance of explanation” (see p. 89), and “the theoretical stance” (see p. 150). Burge calls this view “the simple observation model” of self-knowledge (OESK 252 ff). I think all of these headings risk being misleading, so I’m going to add my own label to the mix and call the relevant model the “self-as-other” model of self-knowledge. This label has the virtue, I think, of precision, which it perhaps pays for by the cost of clunkiness. The entitlement a person has to self-knowledge according to the self-as-other model is of the same sort as that which entitles knowledge of others’ minds; it is not uniquely first-personal. When a person gains self-knowledge in accordance with this model, he treats his mental states and attitudes as mere objects that need not conform to his sense of what is most reasonable. This model is appropriate for explaining the knowledge that Sartre’s akratic gambler, whom Moran discusses at length, has of his gambling habit. This gambler thinks that he never has a good reason to gamble, so were he always to form his intentions in accordance with what he finds most reasonable, he would never intend to gamble. He is akratic, however, precisely because he cannot keep himself from turning towards the gambling tables when he is in their vicinity. He knows this about himself: he knows that, if he is near the gambling tables, he will begin taking steps towards getting himself to those tables to gamble. He knows this about himself not by reflecting on what is reasonable but instead by drawing inferences from his past behavior. To put this another way, he has self-knowledge here not by reflecting, *qua* agent, about what he has reason to do but rather by knowing his habits, which is a sort of knowledge he could have of any other person. This is not a uniquely first-personal manner of acquiring knowledge, and so the

entitlement it involves is neither uniquely first-personal.

Again, a proponent of the rationalist view, like Burge and Moran, need not and should not deny that this sort of self-knowledge is possible, or perhaps even common. What the proponent will deny is that it is the only sort of self-knowledge a person might have, that there is no distinctly first-personal sort of entitlement to self-knowledge. Many people, for many reasons, deny this—that is, many people, for many reasons, think there is no fundamental epistemic difference between the knowledge we have of ourselves and the knowledge we have of others. Today I want to look at one approach that, I think, must hold that there is no fundamental epistemic distinction between self-knowledge and knowledge of other's minds. We might label this approach *the common mechanistic approach to reasoning*. I'm including 'common' in the title because, as my title indicates, I think there may be a more careful, more refined understanding of mechanisms that has room for critical reasoning and for its attendant uniquely first-personal sort of entitlement. Right now, though, I want to discuss the common mechanistic approach to reasoning and why I think, on it, there is no room for this unique sort of entitlement. The view I will take as an exemplar of this approach belongs to David Velleman – let's now turn to that.

## II. Velleman's view of practical reasoning

Velleman develops his account of practical reasoning in the context of his account of agency. He takes the task of an account of agency to be that put forward in the 1970s by Harry Frankfurt: separating what we might call mere agential activity from full-blown action. According to this distinction, actions are the result of agency – they are performed by agents, and they are under the agent's self-control. Mere agential activities are something less. To be clear, they are not involuntary activities, like peristalsis – rather, they are almost actions, except they involve an agent behaving, in some sense, without self-control. The paradigmatic examples are those that figure front and center in Frankfurt's

writing, viz., examples of weakness of will. For an example, consider the person who is trying to quit smoking without much success. Imagine someone who sincerely wants to quit, who knows she'd be healthier if she quit, yet still finds herself giving in to temptation and lighting one up. Lighting a cigarette is, quite obviously, not the same thing as peristalsis, yet after Frankfurt we may label it "mere agential activity," not action, because this person fails to exhibit self-control when she lights up. Were she to refrain from lighting up, were she to, say, refuse a cigarette when it is offered to her, her activity would be more than mere agential activity – it would be full-blown action. To be clear, a committed smoker might smoke agentially – it is not the harmfulness of smoking that makes a given act of smoking less than full-blown action. What makes a given act merely agential is an intrapersonal conflict, one of thinking that something shouldn't be done and then going ahead and doing it anyway.

One reason for focusing on these examples is to demonstrate the inadequacy of what might be called the "standard view" of action. It seems that this "standard view" lacks the resources for distinguishing between mere agential activities and full-blown actions. According to the standard view, an action is performed just in case it is the result of the rational combination of a desire with an appropriate instrumental belief. The sense of 'rational' here is minimal—a combination will be rational if the instrumental belief is true and if it specifies an effective means for satisfying the desire. So, if I am a committed smoker, and I desire to smoke, my belief that if I reach in my bag, I will find a cigarette that I can then smoke will rationally combine with the desire to produce the action of reaching in my bag for a cigarette. What Frankfurt wants to show is that this style of explanation will work just as well to explain why the person who is trying to quit reaches in her bag for a cigarette—given her commitment to quitting, her weak-willed reaching can't be full-blown action, but nevertheless her reaching can be explained by citing exactly the same sort of belief and desire as explains the committed smoker's reaching. If this is right, then the standard model needs to be modified if it is to be able to mark the distinction between mere agential activity and full-blown action—taken by itself, it doesn't

mark the difference.

Like so many others working on practical reasoning since the 1970s, Velleman inherits Frankfurt's conception of what its task is, viz., to find a way to mark the distinction between mere agential activity and full-blown action. Curiously, though, Velleman does not focus on the sorts of cases that interest Frankfurt, which involve weakness of will, when he sets out to draw the activity/action distinction. Instead, we get examples of Freudian slips, of which Velleman thinks it is natural to say a person's desires take over his action but do so without his control. One such example is that of the President of the Lower House of the Austrian Parliament opening a session by declaring, "I declare that this session is closed!" According to both Freud and Velleman, the President did not want to open the session, so when it came time to do so, his desire not to open the session took over his speech and caused him to say 'closed' when he should have said 'open.' Velleman thinks this involves a belief and a desire combining to produce something less than full-blown action. The president desires not to open the session, he believes that if he uses the term 'closed' the session won't open, so he declares the session closed, but this declaration is less than a full-blown action, for it is a verbal slip—it is not what he would have said were he fully in control of his speech act. On these grounds, Velleman claims that the standard belief-desire account will need to be modified if it is to suffice for explaining full-blown action.

Velleman follows Frankfurt and many others who have written on the matter by claiming that the needed modification is to be made by *adding* something to the standard model. To conceptualize this addition, Velleman invites us to engage in the imaginary project of "creature design," in which we imagine ourselves to be godly engineers whose task it is to design a creature with full-blown agency. We don't start from scratch—we already have at our disposal creatures that are capable of motivated activity, like lizards, larks, and perhaps even Labradors. Presumably, these creatures lack the ability to perform full-blown actions—they have desires, and they may even have something like instrumental

beliefs that determine how they act on their desires, but they are not capable of the sort of activity that an agent performs when she wants to smoke a cigarette but resists in order to live up to her commitment of quitting smoking. Our task is to take a creature that has the capacities of a lizard, a lark, or a lab, and we add to it further capacities until we have turned it into a creature capable of full-blown action.

What sort of capacity will make the final difference in converting our creature into an agent Velleman's answer is that we need to add a regulating capacity, which he specifically describes as a mechanism: "you would add practical reason to the design for motivated creatures, and you would add it in the form of a mechanism modifying the motivational forces already at work" (PPR Intro, pp. 11-12). This mechanism takes beliefs and desires as given, but regulates how they combine to produce behavior. This regulatory mechanism will allow our creature to do what the lizard, lark, and Lab cannot, viz., perform full-blown actions. Velleman claims the specific function this mechanism will have, which will allow the creature to perform full-blown actions, will be "[to reinforce] the agent's first-order motives insofar as the latter were perceived as reasons" (p. 14). Velleman's idea, here, is that certain of our creatures lower-level motives, which are the sort of thing our creature shares with lizards and the like, will be reinforced in a way that is only possible for a full-blown agent. Those that are "perceived as reasons" will be reinforced, and behavior produced by motives that have been so reinforced will be full-blown action.

To make sense of what Velleman has in mind here, we need to spell out (a) what he means by "perceiving a motive as a reason," and (b) what a mechanism would need to be like in order for it to perform this function. Let's take these in turn. According to Velleman, a motive is perceived as a reason just in case it is one that the agent can *understand* as a motive, one that *makes sense* to the agent as a motive. So, for example, suppose I'm writing a paper, and suppose that while writing the paper I find myself wanting to have another cup of coffee. What will it take to perceive my desire here as a

reason for having another cup of coffee? Well, given that I'm in the midst of writing a paper and I desire to continue, I might think that it makes sense to desire another cup of coffee – writing requires focus, coffee helps one focus, so, since I'm writing and need to focus, it makes sense that I'd want to drink this thing that will help me focus. Note that what I aim to make sense here is *my* desires – I am the one who wants to drink the coffee – so the understanding, the making-sense that is part of understanding a motive as a reason is a sort of self-understanding. Because I can make sense of myself having this desire, I perceive it as a reason, so if I go on to act on it and pour myself another cup, my activity will amount to full-blown action.

A bit of agential activity is a full-blown action, then, just in case it is performed with self-understanding, that is, if it is activity that makes sense, from the agent's perspective, for the agent to want to do. Of course, just because a creature has a capacity to perform acts with this sort of understanding does not guarantee that it will always do so; mere agential activity is a possibility for creatures equipped with this capacity for self-understanding. To see this, return to the case of the President of the Austrian Parliament. When he says, "I declare this session closed," his speech act is not produced with an understanding of his motive. Put a bit more carefully, while he is speaking he does not understand this motive for saying exactly what he does – he only discovers this motive later, once the words are out of his mouth and he reflects on what he has said. Velleman thinks his account nicely characterizes the connection between this lack of self-understanding and the lack of agency involved in the action. When the President says, "The session is closed," he is not performing a full-blown action, and the reason it is less than full-blown action is that it is produced without the self-understanding that is constitutive of full-blown action.

With this in mind, let us return now to our creature-design task. If Velleman is right that this sort of self-understanding is what is needed for full-blown action, then what the creature we are designing needs is a mechanism whose function it is (a) to ensure that only acts that involve self-

understanding are performed and (b) to inhibit acts that would lack such self-understanding from being performed. Velleman describes this mechanism as a “rational spring” that assesses the “springs of action” and then intervenes in their operations (p. 12). We’ll want this “rational spring” to function sub-agentially—we won’t want to require that our creature be constantly consciously thinking about what it does. That would be an awfully inefficient creature, and it wouldn’t be one that behaves as we do, because we certainly don’t walk around constantly consciously thinking about what we are about to do next. When this sub-agential mechanism functions properly, then, it produces actions that are accompanied by self-understanding without that self-understanding being conceived of explicitly as a goal by the agent. When the mechanism fails, it opens the possibility of actions being performed without self-understanding, like the Parliament president’s speech act.

Have we, by adding this mechanism, succeeded in designing our desired creature? The creature now has a mechanism that inhibits action that it doesn’t understand. It has thus been designed to act only when it has a belief about what it will do; if it lacks such a belief, and if the mechanism is functioning properly, then it is inhibited from acting. Velleman asserts that decision-making has just the same sort of structure: prior to having made a decision about some matter, a person lacks a self-belief about what she will do relative to that matter, and once a decision is made, so too is a belief made about what she will do. Moreover, making a decision results in beliefs about what a person will not do as well: so, if I decide to have a burrito instead of a tamale, I believe that I will eat a burrito, and I also believe that I will not eat a tamale. Velleman claims that, in fact, this is all there is to decision-making: to make a decision is to form a belief about what one will do, like eating a burrito, which thereby rules out doing the relevant alternatives, like eating a tamale. So, Velleman claims, by equipping this creature with a capacity whose well-functioning ensures that the creature’s activity involves self-understanding, we’ve thereby equipped our creature with the capacity to make decisions. Not only is decision-making a hallmark of agency, but by noting in a given instance whether or not a decision has

been made, we can thereby draw the mere activity/action distinction: a creature performs a full-blown action just in case it does what it has decided to do. Thus, Velleman concludes, we've designed the creature we want—we've equipped it with a capacity that has turned it into an agent.

This might strike you as an odd account of decision-making. It strikes me as odd, but I want presently to set thoughts about its strangeness aside. Right now, I want to focus on the notion of self-understanding that features in this account, which, I will now argue, belongs to a self-as-other model of self-knowledge. Velleman's presents self-understanding as if it involved reading or to listening to a monologue of the mind. Think about the way in which the Austrian Parliament President is described—there is a matter of fact about what the President is going to say, and he fails to perform a full-blown action just because he doesn't wait until he knows that he is going to say it before saying it. If only he had waited, he might have known the words that were to come out of his mouth, and given that these words were already there to be known, the knowledge involved here is readily thought of as a sort of self-reading or self-listening. On this model, full-blown linguistic action—and indeed, given that Velleman generalizes from the case, full-blown action more generally—is not the result of authorship, but of rather passing a censor. The mechanism whose proper functioning is necessary for full-blown action does not produce anything itself but rather carefully “reads” or “listens” to the agent's internal “monologue” and constrains him from opening his mouth until he is sufficiently clear on what it is he will say.

As far as I can tell, this is Velleman's model of self-knowledge—what is to be said (or more generally, to be done) is already established; the person then acts knowingly just in case he waits until he is clear about what is to be said (or to be done). If this is correct, then it seems that the capacity that is responsible for agency does not establish what is to be said or done—that is established elsewhere in the mind. This might already suggest that any self-knowledge that is had of what one will do or say fits the self-as-other model of self-knowledge, but I think the point becomes clear when we look at how

Velleman describes overt acts of decision-making, the kind that would appear to involve critical reasoning as Burge and Moran describe it. Return again to the example of me and coffee, altered slightly so that in this case I reflect and decide to have another cup before getting it. When I reflect, what I'm reflecting on, according to Velleman, is whether or not my desire to have coffee makes sense to act on, given who I am and what I'm doing. Now to be sure, this knowledge about what makes sense for me to do can affect my activity in a way that my knowledge of others cannot affect their activity – for my knowledge of others to affect their activity, it would have to be itself acted on, but in the first-person case there is no such need to act, to implement a plan. Still, the *basis* for making the judgment about what would make sense is the same—it compares the details of a situation involving a person, which just happens to be me, to a theory about that person, which, again, just happens to be me, and on that basis a judgment is made regarding what makes sense. So, the *entitlement* to knowledge is the same as in the other-person case – it is not uniquely first-personal. So, the view of self-knowledge is not the rationalist one; it is the sort of entitlement that fits the self-as-other model of self-knowledge.

### III. Velleman's conception of mechanisms

I hope I've adequately made the case that Velleman's view of self-knowledge is the view Moran and Burge are both against generalizing, which I have been calling the self-as-other view. If I haven't convinced you of this, I invite you to say why during the Q&A. Right now, I want to show why I think Velleman adopts the view that he does. I don't think it is an idiosyncrasy on his part—I think instead that Velleman has to take the view he does, given (a) what appears to be his particular conception of mechanisms, and (b) the way in which mechanisms, conceived of in this particular way, have to feature in a creature-design account of agency. Let's take both of these in turn.

It seems to me that Velleman implicitly thinks the following about mechanisms. I say 'seems' and 'implicitly,' because Velleman, like many other philosophers whose writing concerns the nature of

the mind, uses the term ‘mechanism’ without giving any account of what he means by it. Velleman appears to think that characterizing the mind in terms of mechanisms allows for the possibility of reductively explaining its activity in terms of some basic, empirically respectable vocabulary, which is itself an explanatory desideratum. Mechanisms appear to be useful for reductive explanations of this sort in no small part because they can be functionally characterized. To see the attraction, consider a functional characterization of the heart. The heart’s function is to pump blood. Were we to anthropomorphize this function, we might say that the goal of the heart is to pump blood; if we got carried away with the metaphor, we might say that pumping blood is what the heart wants or intends to do. Of course, such talk is metaphorical; the heart is not caused to pump blood by any wants or desires that are internal to it, but rather by things like the activity of the sinoatrial node and the atrioventricular node. Here, I think, we get at – you guessed it – the heart of what I think someone like Velleman finds attractive about functional characterizations as they apply to mechanisms: they explain the appearance of intentional activity, like the heart aiming at the goal of pumping blood, without attributing intentions, desires, or the like to the thing whose activity is being described. The hope for someone like Velleman is that all of the mind’s activity can be explained this way: the appearance of intentionality can be characterized in terms of the functions of mechanisms, but the causal explanations of the activity of these mechanisms will ultimately be given in an empirically respectable vocabulary, e.g, that of biochemistry.

Having said only this much, I think we can already see why the rationalist’s account of critical reasoning is something that someone like Velleman would be reluctant to accept. The present point has nothing to do with self-knowledge; rather, it just has to do with the activity of critical reasoning itself. Burge and Moran both speak of critical reasoning as if it involved a *sui generis* force, one that can be aptly described, I think, as the force of reason. On their view, we at least occasionally form our beliefs and our intentions because we judge certain considerations to be rationally superior than others. If we

do this, then it seems like in these cases, at least, a distinctively rational force determines what we end up believing, intending, or even doing. I'm not sure exactly what Velleman thinks about this, but I would not be surprised if he or someone like him found this occult. Keep in mind that the full impact of the force requires self-consciousness—it is *critical* reasoning we are discussing, reasoning that involves the self-conscious evaluation of the rational credentials of a given belief, intention, desire, or action. Even if Velleman can accept the existence of something that might be called the force of reason, it is clearly, on his view, not a force whose full impact requires self-consciousness, for if there is such a force, it is already operative in the creatures from which we began our design project.

From Velleman's point of view, matters are only made more obscure by the rationalist's insistence that this self-consciousness is a sort of self-knowledge that involves a distinctly first-personal entitlement. One way to represent this apparent obscurity, I think, is to note that if there is such a thing as a mechanism of critical reasoning, then on the rationalist's view the agent's knowledge of its activity is *internal* to the mechanism itself. On Velleman's view, self-knowledge, though essential to agency, is always external to the mechanisms of reasoning. To understand the relevant senses here of internality and externality, let's reflect on the various possibly instances of self-knowledge in a full blown action performed by our creature-designed agent. First, our agential creature may have self-knowledge of the beliefs that enter the reasoning mechanism as input. It needn't have this knowledge, of course, but that is not what is presently relevant; what is relevant is that even if the creature knows what beliefs enter the reasoning mechanism as input, it need not know how, once inputted, they are evaluated as reasons for or against a given claim. To know this, the agent would need to know of the internal workings of the mechanism. Here again, though, the process of reasoning itself is distinct from any knowledge of the process. By a sort of self-monitoring, the agent might "watch" the mechanism do its thing and see it deliberate and conclude, but the proper functioning of the mechanism does not require any such monitoring. The proper functioning of the mechanism of

reasoning itself produces no knowledge—rather, it produces facts, which can be objects of knowledge for some other capacity, or faculty, or mechanism of knowledge. It is in this sense that the knowledge of what goes on in the mechanism—which will include knowledge of the relevant reasons—is external to the mechanism itself. On the creature-design model, it has to be so external. The creature designer must deny that it is internal to the reasoning mechanisms of non-agential creatures—were it so internal, there would be no need to add self-knowledge to the creatures, as Velleman does, to produce agency. The creature designer must also deny that the mechanism that is added to the lower-level creature involves the relevant sort of knowledge as an internal feature—were he to allow this, it would defeat the whole point of the method of creature design. Anyone proceeding as Velleman does must then, it seems, conceive of the self-knowledge involved in critical reasoning as external to the mechanism of reasoning.

Now, maybe this is too quick. Mechanisms characteristically have parts, so maybe the critical reasoning mechanism has one part that does the reasoning, and another that does the epistemic work, producing self-knowledge, but both parts are to be understood as belonging to a single mechanism of critical reasoning. Now note that on this model, the reasoning component of a given act of critical reasoning might go off without a hitch without self-knowledge being produced. In that case, though what happens might not amount to an act of critical reasoning, the reasoning itself will be flawless—the only problem will be the lack of knowledge of that reasoning. If this lack of knowledge constitutes a lack of agency, it is only because it wasn't available for self-regulation, had it been necessary—but the reasoning itself was fine, so it was not necessary. This way of thinking of reasoning, self-knowledge, and agency fits with the specific way Velleman describes the role of self-knowledge in producing full-blooded actions. It is a regulatory mechanism, one that is only needed in case reasoning goes wrong. If a given creature's reasoning mechanism always went right—which, on Velleman's view would amount to its products always fitting the agent's conception of what

they should be—there would be no need for such a mechanism. It is only to keep us from believing, intending, or acting in a way that doesn't make sense to us—if this never happened, there'd be no need to incorporate a knowledge mechanism, or a knowledge part to the overall mechanism. By contrast, according to the rationalist conception, the role of self-knowledge is not regulatory—it is constitutive of critical reasoning, an activity in virtue of which mature reasoners are reasoners, an activity that, should a being not be able to perform it due to lacking the relevant capacity/mechanism, would disqualify the being from counting as fully rational. The idea of mechanisms Velleman is working with cannot and does not give self-knowledge such a constitutive role in reasoning. It might be necessary for full-blown agency, but only as a regulative capacity on capacities of reason that are as rational as they can be in creatures that lack the regulative capacity.

We are now in a position, I think, to see that this aspect of Velleman's view is not idiosyncratic—it is a view to which anyone working with his conception of mechanisms in a creature-design model is committed. If one thinks that mechanistic explanations are good explanations because they hold out the promise of explanatory reduction, and if the base-level description, the one that provides explanatory grounds for the phenomena to be explained, involves no mechanism of reasoning that includes knowledge of its own functioning as an internal feature, then all self-knowledge, even that enjoyed by agents engaged in full-blown action, will be external, i.e., distinct from the mechanism or mechanisms involved in the production of knowledge. To assert that knowledge is internal to the mechanism of reasoning is simply to give up on the creature-design project, if that project is limited to building a creature with mechanisms as they have so far been described.

As I indicated at the outset, I think there might be another understanding of mechanisms that does not rule out the possibility of a mechanism of critical reasoning that, in accordance with the rationalist conception, involves self-knowledge as an internal feature of the mechanism. Before turning to that, however, I want briefly to consider an objection that many, perhaps including many in the

present audience, may have: why think that the rationalist is correct? Why think that critical reasoning involves this uniquely entitled sort of self-knowledge? Why think that Velleman is wrong? In support of this line of objection, one might note, first, that we rarely reason explicitly, and second, that there is no end of psychological experiments that demonstrate contexts in which the reasons we provide for our words and actions. Given all of this, why think that critical reasoning, as the rationalists understand it, is a genuine phenomenon? My response to this will be terse; if it is unsatisfactory, I invite further discussion on the point during Q&A. Here's that response: to deny the existence of the rationalist conception of critical reasoning is to make academic writing and discussion, including philosophical writing and discussion, unintelligible. The goal of such writing and discussion is to support views and to convince others not by brute force, not by manipulation, but by argument, which involves the presentation and evaluation of reasons. To attempt to argue against this thesis appears to me to be a self-defeating enterprise, the kind a continental philosopher might describe as a performative contradiction.

#### IV. The possibility of a mechanism of critical reasoning

Let me proceed now, at last, to take this paper's leading question head-on: is there a mechanism of critical reasoning? To support my answer of, maybe, I'll draw on some remarks from Peter Machamer's 2004 paper "Activities and Causation: The Metaphysics and Epistemology of Mechanisms." Machamer opens this paper by describing the MDC view forwarded by him, Lindley Darden, and Carl Craver. This is a dualist view of mechanisms: it characterizes mechanisms as entities and activities, and it does not seek to reduce one of these to the other. In this paper, Machamer has the following to say about activities. First, they are the producers of change. This is a central claim of the MDC view, and it is one that Jim shines some light on in his 2004 "Synthesizing Activities and Interactions in the Concept of a Mechanism, where he tells us to

understand ‘produce of change’ to mean “a type of cause that makes things up from other things.”

Second, Machamer claims that it is part of the metaphysics of activities that they “do the ruling out” of possibilities. Here’s what I take Machamer to mean in talking about ‘ruling out’: it is due to activities themselves that a mechanistic change proceeds from some state-of-affairs A to some other state-of-affairs B and not some further state-of-affairs C. We can demonstrate this with Jim’s photosynthesis example: the increased energy level of photosystem II excites—this is the activity—some electrons into a specific sort of position change. What this activity of “excitement” rules out is other possibilities: for example, that the electrons don’t change position at all, or that they change position in some other way. Finally, Machamer asserts that describing activities as causes is not to commit oneself to any particular view about causes. In both this paper and his paper with Darden and Craver, Machamer cites Anscombe approvingly as having shown that ‘cause’ is, at base, a generic term. Something is a cause, then, only relative to a specific explanatory context.

Taking these three points together, I think we can characterize the rationalist conception of critical reasoning in mechanistic terms. The producer of change in the mechanism just is the judging a consideration to count as a reason for such-and-such . Such judgments cause their results, which is believing that such-and-such , or wanting such-and-such, or intending to do such-and-such , or in fact doing such-and-such. The assessments of reasons and the judgments made on the basis of these assessments are the activities of the mechanism. When the assessments are performed rationally, this rules out that the subsequent judging and that state or states that result from the judging are irrational, at least with respect to the particular standard of rationality that is operative during the relevant act of reasoning. If this is correct, we can understand critical reasoning as an activity, one that rules out certain possibilities in accordance with something like a law (here, the laws of reason), and one that produces change. If this is correct, it looks like we can entitle ourselves to speak of a mechanism of practical reasoning, at least as far as Machamer’s discussion of mechanisms goes.

In saying that judging is the producer of change, I'm claiming that the judging is the cause of the resulting belief, desire, intention, or action. If the Anscombean point is correct, there is no need to look for a more basic, more natural cause; indeed, if we do, we risk failing to understand the phenomenon, which is making up one's mind on the basis of reasons. When a person does this, according to the rationalist, she thereby knows the rational basis of the judgment that she has made. This is constitutive of the activity; to judge without knowing the rational basis is to perform some different sort of activity. Because it is constitutive of the activity, self-knowledge is internal to the mechanism of critical reasoning. I have argued at length that this claim is not intelligible on Velleman's view, given the way he apparently thinks about mechanisms, but I can see nothing in the view I have just sketched on the basis of Machamer's remarks that precludes the possibility self-knowledge being internal to a mechanism.

If this is right, maybe there is a mechanism of critical reasoning, one that doesn't fit neatly in an overly reductivist view of the mind, but one that nevertheless might have a place in an overarching mechanistic view of the mind. Maybe. I look forward to learning from you about the plausibility of this claim.