Replies to Critics

Richard MORAN

BIBLID [0495-4548 (2007) 22: 58; pp. 53-77]

ABSTRACT: In this article, I respond to the comments of six philosophers on my book *Authority and Estrangement: An Essay on Self-knowledge*. My reply to Josep Corbí mostly concerns the relation between the two modes of self-knowledge I call 'avowal' and 'attribution', and the sense of activity involved in self-knoweldge; in responding to Josep Prades I try to clarify my picture of deliberation and show that it is not 'intellectualist' in an objectionable sense; Komarine Romdenh-Romluc's paper enables me to say some things about the idea of unconscious beliefs, specifically in relation to the phenomenological tradition; the paper by Hilan Bensusan and Manuel de Pinedo helps me to clarify my sense of the relation of the first-person perspective to the specifically normative relation to one's beliefs and other attitudes; and Carla Bagnoli's paper provides an opportunity to explore some connections between the deliberative stance and the notion of recognition in Hegel and in contemporary philosophy.

Keywords: self-knowledge, deliberation, transparency, attribution, avowal, recognition.

First of all, I want to thank the organizers of the Inter-University Workshop on Philosophy and Cognitive Science, held in Valencia of 2006, where earlier versions of some of these papers were first presented. I am grateful for the stimulating discussions from that event, and for the opportunity to reply to some of them in *Theoria*.

Josep Corbí begins his intriguing response to *Authority and Estrangement* (Princeton, 2001; hereafter *'A&E'*) by describing me as still trapped within the Cartesian picture of self-knowledge. At first I was surprised by this. The Cartesian picture of self-knowledge is usually characterized by the thought that introspection is like a form of inner perception, but one which is peculiarly immune to the limitations of perception of the outer world, since supposedly neither error nor ignorance are possible here. On this view whatever is in the mind is necessarily revealed to introspection ('self-intimation') and whatever introspection presents to consciousness will be free of error or distortion. Together these claims add up to the thesis of the mind's transparence to itself, along with a certain picture of the mode or manner of self-knowledge (the 'perceptual model'). There is room for debate as to what extent the historical Descartes subscribed to this package of views, but this is the core of the picture that has come down to us as 'Cartesian'. Hence my surprise at Corbí's description, since much of *A&E* is devoted to working out an account of self-knowledge that denies the claims of infallibility and self-intimation, and which diagnoses what is wrong with the perceptual model of self-knowledge generally. But Corbí has another set of thoughts in mind, as close to criticisms that are sometimes made of Kant as they are associated with Descartes. In this his criticisms are related to the complaint made by Josep Prades, that my account of both self-knowledge and intentional action is excessively rationalistic. I think I see the aspects of my view that give rise to this charge, but I hope I can show that my account only looks that way if we start with an excessively

rationalistic picture of reason itself. Reason comes in many varieties, some more articulate than others. Insofar as it is attentive, discriminating, critical, and directed at a goal, there is reason at work in what Komarine Romdenh-Romluc refers to as "absorbed coping", the skillful pursuit of a practical task, or finding one's way in a complex social situation. In the sense of 'reason' that concerns me, the dancer following the music is also engaging her reason, insofar as there is something she is following, trying to get right, to keep in touch with. As long as there is room for the idea of correction and getting back on track, there is room for the idea that the person has reasons for going this way rather than that way.

Corbí's main line of criticism concerns my characterization of the distinction between theoretical and deliberative attitudes, for he says that it is my understanding of the deliberative attitude that is contaminated with the Cartesian model. This model, however, is not concerned with the thesis of infallibility or the idea of an 'inner eye', but rather with the role of the passions in the person's deliberations and sense of herself. I'm not always certain that I understand this criticism, partly because I say so little about the passions in *A&E* (perhaps *that* is the criticism), so I will have to approach this in stages by responding to the way Corbí sets up his problem. He begins by saying that "under the label 'transparency condition' Moran mixes up a trivial and a non-trivial constraint." In Chapter Two, I define the transparency condition as a norm concerning the relation between two kinds of question a person may ask herself: "The claim then is that a first-person present-tense question about one's belief is answered by reference to (or consideration of) the same reasons as would justify an answer to the corresponding question about the world" (p. 62). Corbí considers the relation between two questions:

(2a) 'Do I believe that *P*?'

(2b) 'Is *P* true?'

And after describing the deliberative attitude toward (2a), such that the person is asking herself this question in the course of trying to making of her mind whether to believe *P*, he says:

> This transparency condition just highlights a *conceptual* connection between (2a) and (2b), when the former is raised in a deliberative manner. And this condition is *trivial* insofar as an agent cannot fail to satisfy it, given that the capacity to adopt a deliberative attitude is constitutive of being an agent. (p. 7, this volume)

While I would agree that the capacity to adopt a deliberative attitude is constitutive of being an agent, actual agents can fall short in various ways of the demands that define them. There seem to be several ways in which the answers to (2a) and (2b) can diverge. Without assuming any irrationality, a person can simply be *wrong* in her belief that *P*, in which case (2a) and (2b) have different answers. A rational believer will know that her believing *P* is not equivalent to the truth of *P* itself, that these represent different facts, one about her as a believer and another one about something else in the world. To fail to recognize this would be a form of solipsism. At the same time, a rational believer knows that her believing *P* is her commitment to the truth of *P*, and

that is why the logic of belief obliges her to conform to the Transparency condition. In the book I try to describe various ways in which an agent's conformity to this condition can be compromised or partial (such cases are perhaps easier to make out in the case of attitudes other than belief, including emotions of various kinds). Finally, even when a question of the first sort is being considered in a deliberative spirit, so that the question the person faces is 'Am I to believe *P*?', or 'Should I regret what happened?', it is not guaranteed by logic that the question is answered as equivalent to 'Is *P* true?' or 'Is what happened regrettable?'. From another perspective attitudes are themselves states of affairs, or conditions of the person, and there may be reasons of more than one kind that recommend having this attitude. It may be irrational or even impossible consciously to adopt a belief for practical reasons, but the fact that being a *P*-believer can be assessed practically as well as epistemically means that the equivalence of questions like (2a) and (2b) is not empty. (On the relation of *kinds* of reason for belief, see Pamela Hieronymi, 'The Wrong Kind of Reason', *Journal of Philosophy* 102, no. 9, Sept. 2005.)

We get closer to the heart of Corbí's criticism in a passage near the middle of his article, which makes it clear that the Cartesianism he has in mind has less to do with the supposed infallibility of introspection, and more to do with whether the passions can be seen as integral to the self or as alien intruders, and the question of the role of 'external' means of altering one's attitudes. I'll need to quote a moderately lengthy passage before making a few responses.

> My problem is that I don't think Moran provides an adequate elucidation of the resources that are available to the deliberative agent in shaping his life. For, in my view, he is still trapped by the Cartesian Model of the self. Despite his claims on the contrary, Moran interprets the deliberative attitude too much in connection to the traditional role ascribed to the effort of the will. The agent must appeal to the effort of the will to carry out his decisions and keep his commitments. The theoretical attitude is concerned, by contrast, with the discovery of passions that somehow the will must counteract. This leads Moran to treat indirect ways of making one's psychological attitudes more responsive to one's deliberation as the result of adopting a theoretical attitude towards oneself. It is like moving one's right arm with one's left hand, instead of raising it directly. I think, however, that there are indirect ways of altering one's psychological states and dispositions which form a part of the deliberative attitude and play a crucial role in the agent's psychic health. (p. 8, this volume)

I have some trouble recognizing my own view in this account, for I don't claim anywhere that the "effort of the will" is involved in ordinary deliberation or in making up one's mind, nor do I conceive of the theoretical attitude as concerned with "the discovery of passions that somehow the will must counteract". Why should the relation of the will to the passions be understood in terms of counteracting them, rather than furthering them, developing them, learning from them? While it is often true that "the agent must appeal to the effort of the will to carry out his decisions and keep his commitments", this will only be true for certain decisions, those that are difficult to carry out or which we are tempted to abandon. Most of our decisions and commitments are not like this, however, and the decision to get a newspaper or leave work early today does not require any appeal to the effort of the will. More importantly, even if some decisions require an effort of the will to carry them out, I don't think

that any effort of the will as such is involved in the ordinary *forming* of decisions or commitments of the sort discussed in my book. For there I am concerned to distinguish a theoretical question which is answered by the *discovery* of a desire or a belief from a deliberative question whose answer is the *formation* of a desire or a belief. In a theoretical context I am seeking to discover what it is that I desire, what it is that has been moving me, whereas in a deliberative context I am trying to figure out what *to* desire, or what to think about something. In this latter kind of case I may be thinking about how to handle a situation with a friend, and come to the conclusion "No, on balance, I don't want to tell him what a reader said about his manuscript." My immediate point here is that *coming* to this conclusion does not require any effort of the will on my part. The conclusion I reach *is* my desire; it is not a conclusion that I now have to figure out some way to implement, some way to *make* this conclusion my desire.

This brings us to the question of indirect means. For me it is characteristic of the immediacy of ordinary self-knowledge that when a deliberative question arrives at an answer ('*This* is what I want, what I think, how I feel...") there is no further thing the person needs to do to make this conclusion her belief or her desire, no steps she needs to take to implement this decision, as she would need to do if her deliberations were about what someone else should think or feel. She is not acting upon herself anymore than she is when she simply raises her arm. This was the point of the comparison with the two ways of moving one's arm that Corbí refers to. And here he raises some questions about the role of such 'indirect means' *within* a deliberative stance that take the discussion further and for which I am grateful. First of all, it was never my intention to deny that such indirect means may be necessary in order to restore psychic health, just as someone trying to quit smoking may need to employ external aids and not simply rely on her decision to quit. But when it is a question of my relation to my own attitudes, the need to employ external means to inculcate some attitude in oneself is symptomatic of the fact that the ordinary route of coming to a new attitude about *X* by reflecting on *X* itself has proved ineffective or unstable. Without something interfering with one's relation to *X* itself, there would be no need to employ external means. In a given situation it may well be part of my general deliberative stance that I realize that in *this* case I need to employ such external means to adjust or stabilize my attitude toward something, and here I take it that I am in agreement with Corbí. But in doing so, my orientation to myself becomes more like my orientation toward another person: in seeking to influence my own thought or behavior here, I have to consider what will *work*, what will be effective, and I have to observe the results of my interventions. It is not that there is anything *wrong* with any of this; it is just that in such a situation the person cannot speak with first-person authority about her attitude because *that* question must wait upon the results of her interventions.

Perhaps I have not been as clear as I should be that not all our engagements with the world (our 'attitudes' in the broadest possible sense) are subject to a deliberative stance in the same way, or perhaps not at all, and therefore not all of our attitudes are ones for which the person has any first-person authority. My aim in the book was to account for how there could be *any* such thing as first-person authority, beginning

with the most straightforward case of simple belief, and using the structure developed there to shed light on those less transparently 'cognitive' areas of life where that structure will apply only with more or less strain. Here I find very sympathetic various things Corbí says about the element of passivity involved in such activities as following a mathematical demonstration, altering one's perceptual beliefs, or following a piece of music. Hence I would agree with him in rejecting the following picture:

> For the Cartesian, passions are just facts about the agent which may favour or hurdle his capacity to approach his own telos. But, in any case, the agent's passions play no role in articulating his telos. They are just facts, like external facts, on which the agent must count in order to fix the most appropriate means to reach his goals. Passions have, then, a merely instrumental relation to the agent's telos. To put it another way, the agent identifies himself with his telos, his true self is the one that pursues that telos. And, within the Cartesian Model, passions do not form a part of his true self, they bear a mere instrumental (quite often disturbing) relation to it. (p. 11, this volume)

In contrast to this, Corbí points to the sense in which some forms of passivity are creative for the self and not simply 'facticities' to be worked around or manipulated. And this is surely right, but it is important to distinguish forms of passivity and forms of responsiveness here. On the one hand there is the passivity of taking a pill and waiting for it to take effect. And on the other hand there is being passively receptive to some passion or affect whose sense is not immediately apparent to conscious reflection, and one may need to allow oneself to be carried along in order to see what sense there is in it. In this latter case, what one is passive to is something with its own normative structure, the way a mathematical demonstration has its own normative structure. In that case 'passivity' means that the norms in question may not reveal themselves to what we might think of as 'deliberation', but require something more like giving oneself over to the activity and attempting to be guided by it. Unlike the case of taking a pill, however, this form of passivity is the flip side of a form of activity, since without the active engagement of forms of attentiveness and absorption the passivity in question has nothing to do with the receptiveness or creativity Corbí is drawing attention to. (I take it that this more or less fits with Corbí's distinction between 'low passivity' and 'receptive passivity'. I say more about the understanding of passivity in such contexts in "Frankfurt on Identification: Ambiguities of Activity in Mental Life", in *Contours of Agency: Essays for Harry Frankfurt*; edited by Sarah Buss and Lee Overton, MIT, 2001.)

Komarine Romdenh-Romluc considers three ways of making sense of the idea of unconscious beliefs and finds difficulties with all of them, and then goes on to suggest that progress here can be made by rejecting the idea that "all beliefs can be analyzed as attitudes to propositions." It is the rejection of this thought which she associates with the phenomenological tradition. I find myself in sympathy with much of what she says, but naturally I have a few questions about how she sets up the problem. So let's begin by considering the three approaches to the problem that she begins by considering and rejecting. First, we might take an unconscious belief to be one that is unconsciously endorsed by the subject. This is not a promising approach if 'endorsing' a proposition involves 'entertaining' it (considering it for approval, as it were). For en-

tertaining a proposition would seem to be a conscious mental activity, and "No sense can be given to the claim that a subject entertains a proposition *un*consciously". A second option for understanding unconscious belief appeals to "two disunified centers of consciousness within the same person." This approach is rejected because it "cannot explain all types of unconscious belief", specifically the rapidly-evolving beliefs (perceptual and otherwise) that are involved in the execution of a complex skill. Finally, a third approach explains unconscious beliefs as components of the mind understood as an information-processing system. Since information-systems need not be conscious, and yet are often credited with states like beliefs, we might in principle understand unconscious beliefs along these lines. However, this approach is rejected as well since it cannot account for how therapy could be possible to restore the belief's accessibility to consciousness. As Romdenh-Romluc puts it,

> However, if my suppressed belief is a component of an information processing system, therapy could not restore my authority over it, because information-processing systems are the wrong kind of things to have this kind of authority over. It follows that suppressed beliefs cannot be understood as components of information-processing systems, without giving up the possibility of therapy. (pp. 19-20, this volume)

It is in response to the difficulties encountered by these various approaches that she then recommends that we consider motor-skills as analyzed by Merleau-Ponty and others, seeing them as pre-reflective beliefs about the world which need not involve conscious reflection to guide the subject through the environment.

There is much that I find myself in sympathy with here, but at the same time I was not always sure whose account this was supposed to be a problem for. I take it that there are several ways in which someone's belief might fail to be conscious. To begin with, there is a great variety among what we commonly call beliefs themselves, and therefore a comparable variety in the modes of self-knowledge that may apply to them. Beliefs can be more or less explicit and determinate, more or less language-dependent, and more or less easy or resistant to conscious awareness. And there will be differences in what counts as conscious awareness. I have beliefs about how one singer's performance compares to another, but it may be difficult for me to put them into words. There are certain perceptual beliefs which may be perfectly accessible to me, but which I am only able to indicate by heavy reliance on demonstrative expressions ("The taste is more like *this*; not so much like *that"*). I have beliefs about how to get to my home from across town which guide me on my way, but if I have to give verbal directions to another person I have to engage not only my access to my beliefs but my ability to put them into words. The beliefs themselves may not be sentence-like at all, and more like a map or general orientation. And there are beliefs which are of the familiar, verbal, type that philosophers are more comfortable taking as examples, but which a person may never have formulated for herself. What are commonly called 'tacit beliefs' are of this sort. There may be every reason to attribute to someone the belief that his car is not edible even if this belief has never occurred to her. This is some of the variety of beliefs we commonly take people to have. Is a person normally self-aware of all such beliefs? I don't think so, and in any case I think we will want to

say that self-knowledge is going to be a different sort of thing in these different cases. In some cases, but not all, we will not want to credit someone with self-knowledge of her belief if she cannot tell us what it is. In other cases, self-knowledge of an essentially non-verbal belief may manifest itself non-verbally as well. In a parallel way, what it is for a belief to be *un*conscious will mean different things in different cases, depending on whether the belief is unconscious because tacit, or because the belief resists verbal formulation, or because it is explicit but repressed, or because it is part of the exercise of a complex skill whose performance does not require explicit reflection (or where explicit reflection may interfere with performance).

For these and other reasons, I don't think it is plausible that we should seek a uniform account of unconscious beliefs. As to the first option discussed by Romdenh-Romluc, I am in full agreement that it will not help to think of unconscious beliefs of any sort as those which are unconsciously endorsed by the subject. The verbal belief that one's car is not edible is not conscious, but not because it is unconsciously endorsed by the subject. We are assuming that a tacit belief like this never occurred to the person in the first place, let alone was something considered for approval (consciously or unconsciously) and then endorsed. And the non-verbal beliefs that orient a person in the performance of some skill may also be unconscious, but not because they have been endorsed unconsciously. Such beliefs are still subject to correction and adjustment in the course of exercising the skill, but this does not require explicit reflection upon them. This brings us to the second option discussed, the idea that unconscious beliefs can be understood by reference to "two disunified centers of consciousness within the same person". A model of this sort has been developed by Donald Davidson in his account of self-deception and other forms of irrationality, and one version of it or another can be found in much psychoanalytic and psychological theory. As the reference to irrationality suggests, models of this sort are suggested by cases of beliefs, desire, or other attitudes which not only fail to be conscious on a given occasion, but which stubbornly *resist* being brought to consciousness, and which are actively repudiated by the person when they are offered as making the best sense of what they say and do and feel. For this reason, I did not understand the nature of the criticism of a model of this sort when Romdenh-Romluc rejects it because "it cannot explain all types of unconscious belief". For the sense in which a set of beliefs (perceptual and otherwise) which is engaged in the interception of a ball in the course of a game of football is unconscious is very different from the sense of 'unconscious' that applies to suppressed beliefs and desires which a person is unable to acknowledge. In cases of this latter sort, it is the resistance to awareness itself that suggests the idea of divisions within the mind, and not the fact of the absence of explicit consciousness by itself. Since I take the model of divisions within the mind to be addressed to a different set of phenomena, I don't see it as a failing that it does not apply to the sense of unconsciousness that applies to the ways in which a person orients herself in the pursuit of a bodily skill. The very notions of 'belief' and 'unconscious' are utterly different in the two cases (and I expect that Romdenh-Romluc would agree with this).

Finally, the third option for understanding unconscious beliefs is described as conceiving them as "components of information-processing systems." Here we might think that the general computational model of the mind would fare better than the previous two options both because such a model is entirely general (and hence should apply to the invocation of belief in *any* context, whether verbal, non-verbal, tacit, suppressed, etc.) and also because such models do not build in consciousness to the very identification of belief itself. However, this model is also found wanting because "this conception of suppressed belief rules out the possibility of therapy." The reason for this is given as follows:

> If my suppressed belief is a component of an information processing system, therapy could not restore my authority over it, because information-processing systems are the wrong kind of things to have this kind of authority over. It follows that suppressed beliefs cannot be understood as components of information-processing systems, without giving up the possibility of therapy. (pp. 19-20, this volume)

We may be understanding the commitments of such an information-processing model differently but I would have thought that it was the very generality of such a model that made an objection of this type inappropriate. That is, I don't mean to assume that there can be no objections to a computational theory of the mind, and such objections may be empirical as much as philosophical. But such a theory is intended as a representation of the ordinary phenomena of, e.g., believing that it's about to rain, or seeing an approaching object and moving to avoid it, or learning the way out of a room. The fact that such a model proceeds at what is called a 'sub-personal' level of description means that it seeks to give an explanation of such phenomena as believing, seeing, or learning, an explanation that does not invoke such terms themselves. This is different from *denying* the phenomena of believing, seeing, or learning, and it is different from claiming that a true account of human life will have no room for such concepts. As I understand it anyway, the information-processing model is a difference in level of description, an explanatory framework, and not equivalent to eliminating reference to what is being explained. And while the processes invoked in such a framework may be as different from the ordinary concepts of believing and learning as the language of physiology or mechanics is different from the language we use to describe someone as running and kicking a ball, in neither case need the sub-personal concepts be seen as replacements for the ordinary ones. The nerves and muscles of a person's body do not themselves run or kick, but they are what make running and kicking possible. Likewise, the adoption of an information-processing model of the mind does not rule out the possibility of therapy or its effects, but is an attempt (successful or not) to account for such phenomena. (Whether such a model is philosophically or empirically defensible is another question.)

In addition, if "information-processing systems are the wrong kind of things to have this kind of authority over", then this makes me wonder about the positive recommendation in terms of shifting attention to the pre-reflective level of bodily coping. For the very reason that the exercise of such skills is pre-reflective it would seem not to be the sort of thing over which a person has anything resembling first-person au-

thority. The pre-reflective awareness that makes it possible for someone to ride a bicycle successfully may well be something about which the person has no special awareness at all. That is, with regard to the exercise of such skills a person often has no idea at all how she is able to do what she does, or even much of an idea just *what* she is doing (e.g., whether she turns by leaning or by turning the handlebars, etc.). These would not seem to be dimensions of life where a person is credited with first-person authority, since even in the normal case, self-knowledge here may well be dependent on self-observation and corrected by the better-placed observations made by other people. Naturally, in various situations this authority lapses in the case of self-knowledge of ordinary attitudes too, but it is ordinarily assumed that when this is so, there is some defect in the rationality of the attitude in question and in *A&E* I tried to present an account of first-person authority that would make sense of its presumed connection to rationality. The case of motor-skills seems different in this regard as well, since in many instances of the exercise of a skill (and not only physical skills, but social skills as well) not only is the person not expected to "know her own mind" with the immediacy and authority associated with beliefs and other attitudes, but self-consciousness itself is just as likely to play an interfering as an enabling role in the exercise of the skill.

So I have some trouble making out how the appeal to the pre-reflective level of motor skills is to help with the philosophical problems of unconscious or suppressed beliefs and the ordinary assumption of first-person authority. I agree that something like the belief that my brother is my parents' favorite child may be pre-reflective in the sense of never having been explicitly formulated and an attitude which expresses itself in patterns of behavior (but also: patterns of thought, association, and feeling). And I can certainly see how a pre-reflective attitude like this may conflict with what the person would consciously avow. In such a case, however, it may well be the avowed belief that needs correcting, not the pre-reflective one (which may express something closer to the truth of the situation). If so, and if the pre-reflective belief is a form of absorbed coping with the world, then it is not clear why there should be any need for it to be accessible to explicit consciousness, anymore than in the case of someone riding a bicycle. Such explicit consciousness may be unavailable for reasons that do not suggest anything amiss in the exercise of the skill itself. By contrast, I would argue, the unconscious beliefs that pose a problem for philosophical accounts of self-knowledge are those for which there is an ordinary assumption of first-person authority, an assumption that is grounded in the relation between the avowability of the attitude and its rationality. These would be primarily those propositional attitudes which are suppressed for one reason or another, where this suppression impairs the integration of the attitude within the rest of the person, rather than those pre-reflective attitudes and orientations whose absence from consciousness is part of their normal good functioning.

Josep Prades has given me much to think about in his challenging paper, and I am sure that there are aspects of his total view that I am not understanding correctly. I am

pleased to find that we are in agreement about a number of issues where much contemporary philosophy of action and practical reason takes a different tack, particularly in his rejection of the Humean-Davidsonian assumption that action explanation must always bottom out in reference to the causal interaction of psychological states, viz. the pairing of a belief and a pro-attitude (or desire in the broad sense). Prades defends a more austerely teleological account of action explanation with which I am broadly in sympathy, but he disagrees with me in claiming that in specifying the reason for some action we are *simply* giving a description of its goal, and not justifying it in any interesting sense. Nor, according to him, need the agent who describes the goal of her action be conceiving of that goal as good or worth pursuing in any sense. 'Content Specifying' (or CS) reasons are just redescriptions of the goal of action and have no (further) normative work to do.

I have some questions to raise about how to understand the explanatory work done by CS reasons if we divorce them from any conception of the goal as something good or worth pursuing, but first I want to address a related point, in connection with a criticism Prades makes near the beginning of his paper. In characterizing my account as "an extremely rationalistic account of intentional action", Prades says the following:

> the particular strategy that Moran follows seems to force him to accept the conclusion that any case of action against the agent's best reasons —or the reasons the agents thinks are her best reasons— is a case in which the agent's self-knowledge is impaired. As a general principle this does not seem to be true. I can calmly and lucidly decide to eat a chocolate cake against my best reasons without any obvious deficiency in my self-knowledge. (p. 25, this volume)

In one sense I have to agree here. If I am asked 'Why did you eat the whole cake?', my reply may be: 'It was delicious'. That answer is a manifestation of self-knowledge, insofar as this answer points to a different reason for the action than would be given by the answer, 'Because I had promised my mother to do so'. But if I give the first answer, then in the context in which we are imagining it this naturally raises a follow-up question: 'Yes, I'm sure it was delicious, but you admit that you also had overwhelming reason *not* to eat it. And now you regret it, as you knew you would. So why did you do *that*?' And to this I may have no answer.

Now "reasons come to an end somewhere", as someone once said, and it need not be seen as a defect of ordinary self-knowledge if the person's answers to the question 'why?' give out at some point. But I was thinking of the connection between self-knowledge and acting in accordance with what one judge's best in a somewhat different way, one that comes out better in thinking about a future-directed case than about reflection on a just completed action. I say that there are two forms of self-knowledge, Attribution and Avowal. Attribution will be based on observation and evidence of the same kind that grounds knowledge of other people. In a given case, self-knowledge through Attribution may be superior to that reached through Avowal; superior in the sense of being better confirmed, a better predictor. But the person herself does not speak with any special authority about knowledge gained through Attribution. Self-knowledge through Avowal, by contrast, is the self-knowledge that has interested philosophers in being both 'immediate', that is, not based on observation or evidence,

and the basis for the special authority that attaches to certain expressions of self-knowledge. The primary questions here are why there should be any difference in knowledge available from the first-person and the third-person points of view, and how it could be that the fact of not being grounded in observation or evidence could somehow *contribute* to the authority of this form of self-knowledge rather than detract from it.

Knowledge of my more immediate intentional actions is normally grounded in my decision to do this or that. These decisions needn't be explicit or discrete, but the way I know where I will be next week is in virtue of my general plans, and not by way of a prediction about what is likely to happen to me (as a prisoner may know where he will be next week). The akratic gambler has decided not to go to the gaming tables tonight, this decision being based on his lively apprehension of disaster awaiting him if he does so. The decision not to go would normally be expressed in the simple declarative sentence "I'm not going, it's not worth it", which would express his knowledge about his future action. But of course he knows himself in another way as well. He knows from experience that his past resolutions on this subject have not counted for anything much. When he comes to his lively apprehension of the disaster awaiting him and confidently declares "I'm not going", no one believes him. And when he considers the matter, he has to admit that they are right to dismiss his words. When he makes this avowal, he's not speaking from knowledge, for what he says is not even true. It is in this sense that I have claimed that the akratic person lacks a certain form of self-knowledge. With regard to gambling, this person has lost the right and the ability to know what he is going to do by deciding on a course of action. (To anticipate, this thought is related to some of Carla Bagnoli's thoughts on authority and the second-person.) Of course he may know in another way what he is really likely to do, based on experience and the evidence of his past behavior, and this will undoubtedly be an important form of knowledge for him to have. But knowledge of one's beliefs, fears, and future actions through avowal no small thing to lose, nor is it something whose loss could be made up for in other ways. It is, I have argued elsewhere in *A*&*E*, the primary way an agent knows such things about herself, and it is intimately connected with the rationality of the actions and attitudes that are known in this way. And what he knows about what he is likely to do through experience does not come with the immediacy or first-person authority that have made this form of self-knowledge of philosophical interest. His friends and family may know much better than he does here.

None of this means that when he is there at the gaming tables laying down his money he doesn't know what he is doing. Nor does it mean that if asked why he is there he cannot speak lyrically and informatively about the genuine attractions, even the genuine satisfactions, of this activity. Akrasia is not like acting in a trance, nor is the person in this state unable to tell us what they are doing or why it seemed compelling, even a good idea. With this reference to a "good idea", however, we come closer to Prades' main line of criticism, which is stated early in the paper in the following terms:

> In my opinion, the Anscombe-Moran principle that intentional action requires a "primary reason" could only be true under a teleological reading of "reason": to specify such a reason is just to specify the purpose, the goal of the action. By itself, this notion of primary reason does not guarantee any kind of relevant justification: it has nothing to do with considering that the goal is "choice worthy" or "good". (p. 26, this volume)

Here I'm pretty sure that we disagree. I think that even a purely teleological sense of reason must answer to some notion of the choice-worthy or good. Prades also does not agree with the claim that desire is connected with finding the object of desire good or choice worthy in some sense. At one point he says that "the obvious fact that a desire is not a simple disposition, does not ground the conclusion that a desire is a disposition for something that is considered as worthy." And it is true that this consideration alone does not show that desire 'aims at' something considered good or worth pursuing. But other well-known considerations do seem to point us in this direction, and I don't see how Prades has addressed them. One set of considerations begins by taking account of the distinction between 'brute' or appetitive desires and 'judgment-sensitive' desires. Brute desires like thirst or fatigue or thirst may be connected with no judgment whatsoever, let alone a judgement of goodness or choice worthiness. A desire of this sort is not a possible conclusion of any deliberation. A judgment-sensitive desire, on the other hand, is of the right logical type to be the possible conclusion of deliberation. In thinking over a trip I plan to take, I can arrive at the *conclusion* that I don't want to take the train after all, that I really want to fly. Now it is true that to claim this difference in logical type between the two types of desire is not yet to say that the object of judgment-sensitive desire has to be conceived in terms of goodness or choice worthiness. But without the appeal to something good or choice worthy about flying over taking the train I don't see how we can make sense of the ordinary deliberative formation of desires like this.

Prades agrees with Anscombe that it belongs to any intentional action that it is subject to certain forms of the question 'why?', and that the answer to such question will elucidate the goal or purpose of the action. However he acknowledges this in a restrictive way:

> The grain of truth behind the idea that intentional action is (normally) action for reasons, is that intentional action accepts certain paradigmatic why-questions, questions that ask the agent to specify the content of the intention-with-which she acts. (p. 28, this volume)

What is noteworthy is that this reference to the applicability of 'Why?' questions does not mention goodness or anything else normative as part of the answer to such a question, and this omission is deliberate. Part of the reason for Prades' balking here may be the thought that Anscombe's thesis is tied to a certain philosophical picture of the kind of goodness in question, a traditional invocation of the True, the Good, and the Beautiful. But I take it that goodness, like reasons themselves, comes in many different varieties and grades of value, and that these are not all commensurable with each other. A desire need not be taken to be aiming at something especially ennobling for it to conform to Anscombe's requirement. What we may think of as perverse desires still conceive of their objects in ways that specify what is found choice worthy in them. I may be ashamed of my desire for something, because I take it not only to be a

desire for something both trivial and disgusting, but more, to be a desire for that thing specifically *because* it is both trivial and disgusting, something I desire *under its aspect* as trivial and disgusting; and yet I still acknowledge it as my own. But this is not so much to deny the connection between desire and some assessment of the characteristics of the object that make it desirable, as rather a reason to broaden and deepen our sense of what such an assessment can be, and how it can provide the coloring and descriptive content of a desire. For surely some activity's disgusting character can be just what appeals to someone and makes the pursuit of it alluring, even rewarding, even if it conflicts with other values of his. What would seem, to me anyway, to leave the character of desire behind is if the object were pursued under no aspect at all, with nothing in its presentation to the agent (consciously or unconsciously) under which it is somehow appealing

Part of what makes me wonder whether a broader notion of 'reason' or 'justification' would bring our positions closer together are passages like the following:

> What does not follow, pace Moran and many other anti-Humean accounts, is that CS reasons are necessarily justificatory in any interesting and practical sense: normally their only justificatory role is to refine the description of the intention with which the agent acts: an intention can be re-described by appealing to certain CS reasons. (p. 28, this volume)

If CS reasons are not justificatory in *any* interesting sense, then there would seem to be no difference between an answer to the 'Why' question of the sort "I did it to save time", on the one hand, and "My hand slipped", on the other. Or else: if CS reasons are nonetheless teleological, specifying the goal of action (i.e., the way that saying 'My hand slipped' does *not* specify a goal), then I don't see how they fail to present justification for the action. In specifying the goal of my action of walking across the room ("to close the window") I do two things. I make sense of my walking across the room. I show that it didn't just happen, but that there was a point to it. And secondly, in specifying my CS reason ("to close the window") I invite the further question "And what is the point of *that*?" or "What is the *good* of it?". We may be disagreeing about what counts as justification in an interesting sense, since to my mind an answer to this question such as "It was getting cold" or even "I just felt like it" count as justifications, and serve to characterize the goal as in some sense worth pursuing. In daily life our justifications, like our goals, are not always terribly interesting, but the philosophical point is to distinguish the different forms of explanation, and to characterize the specific demands that are made on the understanding of an event when we see it as a human action.

At the same time, something that I think keeps our positions apart is the fact that I don't see Prades' account as trying to display the coordination between the reasons that *explain* someone's action and the reasons the agent himself considers when she is considering whether to pursue one course of action rather than another. At one point Prades says that "the reasons for which an agent acts (insofar as normal intentional action requires such reasons) do not have the role of justifying the agent when she makes her mind up. For those reasons are just the content of her decision", and here I have to confess to not understanding the claim. For when the agent is making up her

mind and deciding how to act, how can the reasons she arrives at *not* present themselves to her as justifying reasons? If she is making up her mind about what to do, how could she not be engaged in looking for reasons that would justify one course of action over another? To insist on this much seems to me perfectly consistent with thinking of reasons for action as 'CS reasons' in the sense specified by Prades. The answer to the 'Why question' need not itself mention either desires themselves or anything normative for the content of the answer to be functioning as a justifying reason for the action in question. ("because tomorrow is the last day of the Caravaggio exhibition"). But at the same time, what is presented as the CS reason still has to present itself to the deliberating agent as something it makes sense to pursue (in the broadest sense). For consider the response if the agent is told that she is mistaken, that tomorrow is *not* the last day of the Caravaggio exhibition, that in fact the show closed last week. If this doesn't make a difference to the agent, then I can't see how it was functioning as a CS reason in the first place. And it seems clear that the reason this news must make a difference to the agent is that she has now lost her stated reason for finding the trip to Barcelona to be a good or a sensible idea.

Toward the end of the paper, Prades responds to this sort of consideration as follows:

> It cannot survive my abandonment of the plan of living in Paris, when I learn that in fact I will not receive the expected funding. All right. Nevertheless, this is just a particular case of what I described before as CS reasons: to say that I want to learn French is an incomplete description of my attitude that is refined when I say that I want to learn French as my chosen way of having a happier life in Paris. I want to go to the station now, because I want to go to Barcelona: my going to the station is only a part of my intention. The intention is better specified by describing it as my chosen way of going to Barcelona. (p. 33, this volume)

In such a case we would certainly need an explanation of what is meant by saying that the intention is "better specified" when the description comes to include something seen as good or desirable by the agent. "Better specified" cannot simply mean 'more precise' or 'more predictive'. Further, when Prades says that his goal of learning French "cannot survive my abandonment of the plan of living in Paris, when I learn that in fact I will not receive the expected funding", we need an explanation of *why* this goal cannot survive this change of plans. Ordinarily (so I would claim) we would explain the fact that the goal of learning French cannot survive this change of plans because it was the integration of this goal within this larger plan that made it seem a good or worthwhile thing to do. With the loss of that integration, the goal of learning French no longer seems (as) worthwhile to him, and *that* is why this goal cannot survive the change of plans. Prades, however, cannot avail himself of any such explanation for why the original goal cannot survive the abandonment of the plan to live in Paris, which leaves me wondering what other sort of explanation he could have in mind.

As I said earlier, my confession of failure to understand is not simply the standard philosophical gesture in these contexts. I am quite sure that there is more to Prades' total view than I have been able to profit from so far, and some of my questions express my certainty that I am not understanding his account correctly. But I am im-

pressed with the radical re-orientation of the understanding of action and practical reason that he is recommending here and in his larger work on the subject, and I look forward to learning more from it.

The title of the paper by Hilan Bensusan and Manuel de Pinedo points to some questions they raise toward the end of their paper, about the role of justification in our thinking about our beliefs, and whether this role belongs primarily with the first-person or third-person point of view. In this they are suggesting a possible parallel between a thought of Bernard Williams' to the effect that certain categories of moral thought belong to the external evaluation of an agent or an action, and have only problematic application within the thinking of the agent herself. They suggest a parallel claim with respect to the role of *epistemic* categories of appraisal, and suggests that there is a similar problem with their having a central guiding role in the thinking of the epistemic agent herself.

Much of the argument concerns the understanding of 'transparency' as that idea appears in the account of self-knowledge in *A*&*E*, so I will begin there. Drawing on a theme in Wittgenstein, which is taken up later by Roy Edgley and then Gareth Evans, I define 'transparency' as a relation between two sets of questions; a question about oneself ("What is it that I believe here?") and a question about the world ("What is in fact the case out there?"). In the passage from p. 62 that I quote in my response to Josep Corbí, I define the 'transparency' of one question to another one as the relation between them such that one question (about my own belief) can be answered by consideration of the very same reasons that would justify an answer to the world-directed question. I then go on to claim that it is a norm of belief that a person should be able to answer the question of what he believes about something (i.e., a question of the first sort) by consideration of the reasons relevant to answering the parallel question about the facts themselves. I go into this here because at various important places in their paper, Bensusan and de Pinedo are employing a sense of 'transparency' quite different from this, and this makes it difficult to tell to what extent their conclusions are to be derived from the sort of account I am presenting. Near the beginning of their paper there is the following:

> Because my beliefs are transparent, I can avow what I believe with no appeal either to anybody's behaviour or to my internal makeup. My own knowledge of my beliefs has this special channel of access that involves the transparency of the world to me; of course, I can find out about my beliefs in much the same manner I use to discover what other people believe but transparent access to the world is an alternative, first-personal road that takes me only to my own beliefs. (p. 35, this volume)

In this passage there is reference to the transparency of one's *beliefs* ("Because my beliefs are transparent"), as well as reference to "the transparency of the world to me". I think these are ideas that have their home in epistemic theories of a different sort (such as arguments about direct realism), and they do not have straightforward application to the issues I am dealing with in *A*&*E*. (But I will try to say more about this in a moment.) In any case, the way I am employing the notion of 'transparency' in my book only makes sense as a relation between sets of *questions*, and as such *this* notion

of transparency does not have any clear application of the idea to my relation to my beliefs themselves or my relation to the world. Perhaps when they speak of 'my beliefs being transparent' this should just be taken to mean that I know them with a special immediacy. But if so, then this threatens to confuse the strategy of explanation that I am attempting in the book, for the idea that my beliefs are known to me 'immediately' is one of the primary phenomena that I am trying to *explain*, and specifically to explain by reference to the norm of transparency of one set of questions to another set. Hence if 'transparency' meant the same thing in both contexts my account would be no explanation at all, since it would only be re-stating the phenomena to be explained. This difference between our understandings of the term 'transparency' will also be relevant when we consider their idea of beliefs which are "not transparent enough".

First, a word about "double access" and normativity. Much of the argument of my book is taken up with the claim that there is a fundamental difference between knowing an attitude of mine because I can *attribute* it to myself (on the basis of behavior or other evidence) and knowing an attitude of mine because I can *avow* it, that is, overtly express it as mine without reliance on behavioral or other evidence. It is self-knowledge through avowal, I claim, that is the form of self-knowledge that has attracted philosophical interest, and which exhibits a difference between first and third-person cases. A question I press across a couple of chapters is: why should the normal route of self-knowledge through avowal *matter*, either to rationality or to what we think of as ordinary self-knowledge? Why wouldn't a form of attribution that was just as immediate and reliable be just as good? It is in pursuing this question, I claim, that we begin to see how the immediacy of ordinary self-knowledge is related to something deserving the name of first-person *authority*. It is this question of the relation and priority of two modes of knowledge of oneself that Bensusan and de Pinedo are referring to when they speak of "double access", and they understand the possibility of such "double access" to be the source of the normative responsiveness of beliefs:

> In fact, double access to beliefs is what makes them responsive to norms. It makes beliefs corrigible on the light of other people's beliefs. If we had no more than third-personal access to our own beliefs, we would be unable to attribute to them any capacity to guide action and thinking: my own beliefs would be oblivious to my mental life —or to anything that is in some sense internal to me, for that matter. (p. 37, this volume)

There is certainly a sense in which the possibility of knowing what I think both through avowing it and through attributing some attitude to myself is responsible for the way beliefs are responsive to norms, but I think it's a restricted sense. That is, my beliefs can be responsive to norms in the sense of explicitly responding to criticism and counter-evidence presented to me in the course of argument with another person. Explicit reasoning and justifying are possible for creatures who can avow their beliefs, because they can conclude to them, or decide on them, in the course of deliberation. And responding to norms in the course of actual argument with another person is possible because each of the parties is understood to be in a position to speak *for* herself and not only *about* herself. And a person would not be in a position to make a *claim*, giving her word on something, and hence invoke responsiveness to norms in

that sense, if her only access to her beliefs were through attribution. But I also think that this is not all there is to the idea of belief's responsiveness to norms, since I would also want to say that, for instance, the flow and adjustment of ordinary perceptual beliefs is norm-guided and indeed responsive to norms. That is, it isn't only that norms *apply* to beliefs which are not subject to explicit deliberation (e.g., perceptual beliefs, animal beliefs), but that the norms we would apply to such beliefs are also part of what *explain* the normal good functioning of the beliefs in question. If we are comfortable talking about animal beliefs at all (and I for one can't see how this is dispensable), then surely their beliefs must be said to guide action (as well as other beliefs), even though such creatures have neither the capacity for avowal *or* the attribution of beliefs to themselves (lacking these concepts entirely). In that case, the lack of "double access" does not mean that the cat stalking the bird is not guided by a belief about the bird's location, or indeed responsive to new evidence of the bird's suddenly taking flight. It's true that for a mature human, being a believer means, among other things, being able to say what you think, which is a matter of avowal, and being able to give your reasons, both of which require the recognition of a certain authority on the part of the speaker. But at the same time, this concept of belief is a development from a wider notion of belief and other attitudes, along with their responsiveness to norms, that applies to creatures, both human and non-human, without any of these capacities.

This brings me to the final section of Bensusan and de Pinedo's paper, which concerns the question of the proper role of consideration of epistemic norms and virtues in the thinking of the deliberative agent in the course of making up her mind. The thought they are following out here is that there is an epistemic parallel to Bernard Williams' claim that some moral thinking in terms of specific moral virtues or traits involves importing a kind of external point of view on oneself and one's character, at the expense of the more properly first-person perspective directed outward on the situation and what it demands of one. Hence, one's epistemic reasoning as well as one's moral deliberation may be distorted in parallel fashion, and thus fail to be "first-personal enough."

> We claim that, if Williams is right about moral judgments, we can apply the idea that mental content is sometimes not as first-personal as it should be to beliefs in general. We can start out by considering epistemic (or doxastic) virtues, instead of moral virtues. Consider a case of someone that, in the process of acquiring and managing her beliefs, pays excessive attention at how reliable (or empirically adequate, or coherent, or widely accepted) her beliefs are when considered from a third-personal point of view. The suspicion is that she can be misdirecting her capacity to have a third-personal access to her beliefs. (p. 40, this volume)

I'm not sure just what the parallel would be here. I can imagine an emphasis on justification that would be less a matter of concern with the truth and more of a matter of placing oneself epistemically beyond reproach, and perhaps this is related to what John McDowell means by a notion of justification that is more than mere exculpation. Or I can imagine a kind of concern with justification that is part of the selective skepticism that can be employed in the exercise of bad faith and self-deception. That is, in order to avoid drawing some obvious conclusion that is nonetheless disturbing to me, I tell myself that evidence is always ambiguous and that even the evi-

dence of the senses has been known to be misleading, etc., and in this way I keep in suspense the conclusion I wish to avoid, all in the guise of an exaggerated epistemic scrupulousness. And to move closer to the parallel with virtues of character, I can imagine such a person as being more concerned with the appearance (to himself) of epistemic responsibility than with arriving at the truth. But cases like these may be said to express bad faith because they involve a *sham* concern with justification, a sham concern which seeks to defeat the point of justification itself.

If this is right, however, this would be something different from what Bensusan and de Pinedo later describe in the following way:

> In our case just above, however, the person who pays excessive attention to the epistemic qualities of her beliefs can be neglecting the transparency that is open to her as a resource to establish her own beliefs. She will then be paying too much attention to the standards of evaluation for beliefs (that maybe she is ready to recommend and further to maintain) to an extent that would neglect her capacity to examine the world from the perspective of her beliefs; in this sense she can end up holding beliefs that are not first-personal enough —she could have a measure of what we can describe as a case of epistemic bad faith. (p. 40, this volume)

For here again I think we are employing different notions of 'transparency'. To begin with, I'm not sure what is meant by "examin[ing] the world from the perspective of her beliefs", since understood in one way there would seem to be no *other* possible perspective from which from which to examine the world, but if this is right then there is really no sense to the idea of holding one's beliefs side by side with the world and comparing the two. Secondly, when they refer to this subject as "neglecting the transparency that is open to her as a resource to establish her own beliefs", I cannot recognize the notion of transparency from *A*&*E* since that notion concerns a way of answering a question about one's own belief, and not an alternative to a concern with the justification of one's beliefs. The distortion they have in mind in the examples is described in terms of misguided attention to standards of evaluation for belief at the expense of attention to the world itself. But when properly conducted, attention to and criticism of reasons, justification, and standards of evaluation just is the informed and sensitive attention to the world. They are not opposed to each other, and in fact I would argue that the notion of 'attention to the world' itself has no sense apart from notions of criticism, justification, and standards of evaluation.

For these reasons, I think we do not want to draw the conclusion that "as far as the acquisition of beliefs is concerned, the talk of justification (or of epistemic or doxastic virtues) belong in a third-personal perspective", since for the reasons I've outlined I don't see that it would make sense to restrict the application of notions of justification to the third-person perspective. And indeed, the idea that the justificatory norms for belief "belong in a third-person perspective" would seem to be in conflict with the previous thought that "double access to beliefs is what makes them responsive to norms." For there the thought was that it is the *first-person* stance of avowal that makes possible the articulated notion of responsiveness to norms that goes with giving and asking for reasons. I think this idea is closer to the truth, even though I did raise some questions about scope of this claim, assuming the extension of beliefs and

their norms to, e.g., children and other creatures not (yet) initiated into the practices of justification.

I am very grateful to Carla Bagnoli for forcing me to re-think some of the basic formulations of *A*&*E*, but especially for carrying the discussion further from the first-person/third-person relations and asymmetries I discuss there, to the investigation of the second-person stance (something until recently absent from such discussions and which is therefore sometimes referred to as the "missing person"). Any discussion of self-knowledge like mine, which takes an interest in the phenomena of alienation, has to consider the extent to which relations to oneself can approximate to relations to a genuine Other, and to what extent this is not possible. My own thinking here is somewhat complex and probably not explicit enough in the book, even though that question of the irreducibility of the different possible stances toward oneself and toward another was one of the primary motivations for writing the book. So I very much welcome the opportunity to relate my thoughts in *A*&*E* to the concept of recognition, even as I will want to resist some of the implications of Bagnoli's provocative claim that "the authority of self-reflection (and of reason) is best understood as a relation of mutual recognition between self and others, hence from a second-person stance."

First, however, I need make a few interpretive points to make sure we are on the same page here, since on a couple of important points I could not be sure if Bagnoli was presenting some line of argument as representing my own view rather than a view I am discussing which I then distance myself from. One such place is near the beginning of her paper where she is discussing the role of reflection and what she calls the "argument from suspension". After citing p. 144 of *A*&*E*, she says, "When I reflect on my anger, I am free to choose whether to act on it or not. Reflection shows that the occurring mental state does not dominate me." But of course nothing like this is guaranteed by logic alone, nor by the fact that one is a rational agent, for the only rational agents we are familiar with are the finite, flawed and compromised creatures that we are. Hence the anger may continue to dominate me, its psychological force undiminished, even though I thoroughly disapprove of it and may be ashamed of it. Reflection may succeed in bringing me to separate myself from my anger, so that I don't act on it, even while I continue to feel it. But I might also succeed in separating myself from it, and disapprove of it, and yet still act on it in the heat of the moment. Reflection may succeed in bringing me to see that my anger is ugly, unworthy, even thoroughly unjustified, and yet this separation may fail in another, more intimate way. For my rejection of my anger may be thoroughly genuine and sincere, and yet the anger itself continues to structure my orientation toward this person or this situation, directing my thoughts along the channels laid out by my anger. That is, in these cases not only does reflection fail to prevent me from acting on my anger, but it also fails to undo the thinking and feeling, as well as the motivational pathways, that are constitutive of the anger itself. Both one's thought as well as one's action can remain dominated by an attitude that one rejects, and one may be perfectly conscious of this. My

own claim instead is that reflection on one's attitudes will *normally* be part of the deliberation that alters and adjusts them, otherwise we couldn't see attitudes in general as rational. In itself this is not a terribly new idea. What I hope is more novel is my attempt to relate this thought to the asymmetries of Self and Other, and specifically to the fact that a certain central form of self-knowledge can proceed without evidence, and that *this* is explained by reference to the Transparency condition, which is itself to be understood as a consequence of the rational agent's normal ability to constitute his mind through reflection.

Another point takes us closer to Bagnoli's introduction of the second-person standpoint. At times in her exposition she understands my distinction between explanatory reasons and justifying reasons to line up with, or perhaps even be equivalent to, the distinction between third-person and first-person perspectives. Bringing the perspective of justifying reasons to bear in both third-person and second-person contexts is then in preparation for seeing the deliberative perspective as itself essentially second-personal, an internal dialogue with an Other. When I first introduce the idea of 'deliberative reflection', in Chapter Two, I distinguish it from the (merely) normative appraisal of one's state, for this sort of appraisal (e.g., as justified or unjustified) applies just as well to thinking about *another* persons attitudes as it does to my relation to my own.

> The idea of 'deliberative' reflection about one's response is meant to denote something more than simply the normative appraisal of it, the sort of reflection that would terminate in some settled assessment of it. For the mere appraisal of one's attitudes, however normative, would apply equally well to past as well as to current attitudes, and indeed may have just the same application to another person as to oneself. In itself, such assessment is not an essentially first-person affair. Rather, 'deliberative' reflection as intended here is of the same family of thought as practical reflection, which does not conclude with a normative judgement *about* what would be best to do, but with the formation of an actual intention *to do* something. (*A*&*E*, p. 59)

At times, however, Bagnoli seems to take me to be see the standpoint of justificatory reasons (as opposed to explanatory reasons) as exclusively first-personal and having no application to the actions or attitudes of other people at all.

> Justifying reasons are reasons that I endorse as I deliberate about what to do, what to feel or what to believe. This kind of reasons is not available in the third-person perspective, when I consider the matter as a spectator. In the third-personal perspective, Moran argues, reasons can only be explanatory: they amount to rationalizations of why the agent acted in a certain manner. (p. 47, this volume)

I'm not sure I'm reading her correctly here, but in this passage she seems to take herself to be paraphrasing my own view, since she corrects me a paragraph later, by referring to Nagel on impersonal reasons.

> The third-person perspective is not always cast so as to renounce justification. On the contrary, it is often proposed as a theory of justifying reasons for action (Nagel 1970). [...] I call attention onto this alternative construal of the third-person approach to show that the third-person perspective does not necessarily map onto the theoretical or contemplative perspective, nor does it coincide with the domain of explanatory reasons, as Moran's discussion assumes. (p. 47, this volume)

I very much agree with the gist of this, since it is important to my understanding of the deliberative perspective, whose conclusion is a form of conviction and hence brings me to a new state of mind, to distinguish this from the perspective of evaluation or justificatory reasons which have application equally to my own attitudes as well as the attitudes of others, where my own estimation of their justificatory status may well remain a mere appraisal and make no difference to the attitudes themselves. Hence the stance of justification and the stance of explanation both have ordinary application in our third-personal relations to others as well as to ourselves.[1]

The deliberative stance involves making up one's mind and not only evaluating it. And here Bagnoli raises intriguing questions about a place for *this* stance in the context of deliberating not *about* but *with* another person; thus, the second-person standpoint. There are several important strands to this thought, and I will only be able to touch on a few of them here. One such strand concerns the notion of recognition, particularly in the Fichtean/Hegelian guise of *Anerkennung*. Bagnoli suggests transferring something of this structure of normativity to the understanding of the first-person situation of deliberation, at one point saying "The normative capacity to attribute or withdraw recognition to a given mental state [anger] is where the mind resides." I think we need to tread carefully here, since the concept of 'recognition' is ambiguous in multiple ways, some of which allow for equivocation between the Hegelian motif of *Anerkennung* (to which the prospect of mutuality is internal), and the stances of endorsement and identification with an attitude associated with the writings of Harry Frankfurt.

As I understand it, the Fichtean/Hegelian notion of *Anerkennung* concerns the relation of one self-consciousness to another, as given a kind of definitive formulation in Hegel's dialectic of recognition between Lord and Bondsman. One seeks the recognition of another, a self-consciousness which is conceived of as free to give or withhold that recognition. In fact, it is only *as* free that the response of the Other could have the value or meaning of recognition as such. This is why it is a self-defeating strategy within this dialectic for one party to seek the total control or objectification of the Other as a way to secure the recognition it seeks, for with this annulment of the freedom of the Other is also lost the possibility of any response that could count as rec-

[1] Another place where I am somewhat more explicit about this is p. 130:

> "From the agent's perspective, the question of the truth of his beliefs is prior to the question of how they will dispose him to act. Beliefs 'aim at truth', and do not enter into his practical reasoning in a way that brackets the question of their truth. The interpreter, on the other hand, will be interested in how beliefs explain behavior, and this is a role played by false beliefs quite as often as by true ones. Any representational state will have such a dual aspect, one under which it is transparent to the world in a certain way, another under which it makes a contribution to the behavior of the agent. Naturally these different interests in belief are not *restricted* to the first- and third-person uses, respectively. In communicating and reasoning with others, for instance, we are concerned with the truth, and not just the explanatory adequacy, of the beliefs we take them to have; while, on the other hand, in understanding oneself, one will sometimes need to bracket the question of the truth of one's beliefs and concentrate on their explanatory role."

ognition. 'I' as a self-consciousness seek recognition from another, and the lesson of the dialectic is that this cannot be pursued unilaterally, that even pure self-assertion in this arena must recognize itself as inextricably dependent on a freedom that is not its own.

This concept of recognition is implicated with various related concepts, such as validation and respect, as well as endorsement and identification as these ideas are at work, e.g., in the writings of Frankfurt and others. Frankfurt, famously, identifies freedom of the will with the capacity to form 'second-order volitions', that is, the ability to have the will one wants to have. A desire or other attitude may move me and yet be one that I thoroughly disapprove of, and perhaps seek, whether successfully or not, to rid myself of. In this situation I am said to withhold endorsement from that desire. I may see clearly that it is part of me, or operative *in* me, but I am alienated from it and do not identify with it. So there is a certain resonance with the Fichtean / Hegelian notion of recognition. In a Frankfurtian scenario, when I do not identify with a certain desire that is in another sense still 'mine', I may be said to invalidate it, even if its force remains undiminished by this ruling. I withhold recognition from this desire not only in the simple sense of disapproving of it, but also in the sense of denying it a certain *standing* within my general deliberations. My desire to return to the gaming tables, or the excitement that this desire makes vivid for me, does not count for me as a *reason* to go out gambling again, but is instead treated as an intruder, a force to be reckoned with as potentially affecting my deliberations from without.

We may, then, speak of the refusal to identify with a particular desire or other attitude as a refusal of recognition, even a kind of denial of a kind of status to that desire within one's general deliberative household. But the differences between these two notions are as profound as their similarities, for there is nothing here corresponding to the desire for recognition from another self-consciousness, or the corresponding dependence on a freedom that is genuinely Other to oneself. True, the language of 'alienation' in the Frankfurt scenario suggests a confrontation with something conceived of as essentially 'other' to oneself, but this 'other' will be something like the compulsion to gamble or an eruption of anger; that is, *not* another self-consciousness, not something *from* which it would make sense for a person to seek recognition. Hence, I would say, the very idea of 'mutuality' has no application here. There is, however, another dimension of the second-personal stance that Bagnoli has in mind as characterizing, or perhaps even grounding, normativity and justification, and it comes out in passages like the following:

> [T]he structure of justification is also second-personal. Reasons are considerations offered to another to justify a course of action or a mental state. My proposal is that we understand the second-personal structure according to a dialogical model. "I shall" is the conclusion of a dialogue that emerges from the recognition that I should account for my actions to others, and that I should demand justifications from them. I put myself under the rational scrutiny of others and demand the same. What I offer as a reason must count as a reason also for others, and what others offer as a reason must be something intelligible to me as a reason. (p. 50, this volume)

There is much in this that I am in sympathy with, and that I think is very much worth further investigation (some of which Bagnoli herself is carrying out in other pa-

pers), but I still want to enter a couple of reservations. It is one thing to claim that reasons must be public, the sort of thing that can be made intelligible to another person, shared or respected by another person, and this is an idea that has roots both in Kantian and post-Kantian philosophy, as well as in Wittgenstein. It is quite another thing, however, to claim that reflection itself involves a second-person stance, a dialogue with a genuine other. For a second-person stance means a stance toward a separate freedom, a "self-originating source of claims" (Rawls), something calling upon me and demanding my respect. I do not stand in such a relation to my own attitudes or to myself. My understanding of self-other asymmetries is grounded in the sense that they are the source not only of the 'privileging' of the first-person in first-person authority, but also the source of various forms of *dis*-privileging of the first-person, the blindspots that are the obverse of first-person authority, but also the fact that there is a range of stances that are reserved for our relations to others and which can only with various degrees of incoherence be adopted toward oneself. Gratitude, envy, forgiveness, promising, and believing a speaker are all, I would argue, possibilities reserved for our relations toward a fully independent other, a separate consciousness and a locus of freedom that is not one's own. Just how to characterize this separateness and independence, and the concomitant dependence on the freedom of another involved in, say, contract or bargaining, are deep and difficult issues, as Bagnoli's reference to Kant on the binding of oneself reminds us.

In insisting on a range of essentially 'other-dependent' attitudes, I don't mean to exaggerate anymore than I do in the realm of first-person authority, but only to insist that the way forward on these problems has to respect asymmetries of both kinds. Let the following example suffice for now. Respecting another person's wishes means giving them weight in my thinking and planning, even when I myself don't see the point of those desires, even when I think them mistaken, even when I may feel that the other person would renounce them if she knew everything that I know about them. In a situation of actual dialogue, I may be unable to bring her to see the pointlessness of what she desires in this situation. And yet a concern for her autonomy may tell me that I must still respect her desires in this case nonetheless, even if I think she would regret the outcome. If there is anything to the asymmetries between self and other that I have tried to delineate, then *this* sense of 'respecting someone's wishes' does not make sense as an attitude toward oneself. A desire that exerts a certain force on me, but which I cannot see the sense of is *not* one I am obliged to respect. (Of course, I may *find* reason to respect it, if I feel that this obscure prompting may well lead me somewhere new, toward something valuable I could not have found otherwise. See my reply to Josep Corbí.) But a desire or other prompting which I take to be mistaken, or exaggerated, or unworthy, or something I will soon regret, is not something I have reason to retain and accommodate within the rest of my thinking and planning. Rather, it would seem that respecting *oneself* here means assuming responsibility for either the alteration or control of this attitude, or in any case that they "cannot enter into my thinking as the fixed beliefs and desires of some person or other, who happens to be me" (*A*&*E*, p. 164).

When Bagnoli says that "the second-personal stance understands rational freedom not as a metaphysical property of the self, but as a practical relation structured by mutual recognition.", there is a sense in which I can see this as continuous with a line of thought that I pursue in thinking about first-person authority as well as the authority to declare oneself in speech. This is the sense in which the declaration of an intention, a belief, or a vow is something that requires the recognition of a certain authority. Before a child is a certain age, we do not accept promises from her, and before a certain age, even something so simple as a declaration of intention will be acknowledged only in a qualified manner. First-person authority can be seen as the authority to bring a certain organization to the complexity of one's psychic life, to bring one's orientation toward something (e.g., my immediate future action) to some kind of determinacy. As in other cases this authority exists in being recognized, and in the discussions of *akrasia* we have seen how this authority can be impaired or lost, and the recognition of it withheld. So I suspect my thought here is not so different from what Hegel meant when he famously said, "Self-consciousness exists in and for itself when, and by the fact that, it exists for another; that is, it exists only in being acknowledged."

In closing, I'll try another way of expressing my reservations, and my sense that ineluctable self-other asymmetries are among the starting points for thinking about practical reason, and that this means it will be part of the understanding of what can be normatively binding in practical reason that there are limits to how much of this structure can be housed in a single self, limits to the extent to which one can make normative claims upon oneself like those of a genuine Other. Toward the end of her paper, Bagnoli quotes a famous and provocative formulation of Kant's in the context of the problem of self-binding:

> When describing the working of "conscience", which is one paradigmatic exercise of self-reflection, Kant notices that the "A man who accuses and judges himself in conscience must think of a dual personality in himself, a doubled (*doppelte*) self…" (Kant 1797: 438)

The accusing and judging conscience is a familiar character in philosophy, literature, and life itself. But in thinking about self-accusation and self-judgment I think we should make more of the fact that we don't have anything like the same ease in talking about this agent who can accuse us (who is oneself) as also being in a position to do such things as *falsely* accuse, and therefore apologize, retract the accusation, and make reparations and ask for forgiveness from the injured party (ourselves), be punished or removed from office, or issue a blanket pardon (from oneself to oneself). Judging, even condemning oneself has a much more secure place in our discourse than do parallel normative stances toward oneself like pardoning oneself, waiving rights of complaint against oneself, releasing oneself from a promise. Various aspects of a relation to an existing Other to whom we are obliged in one way or another can only be described as obtaining in one's relation to oneself with varying degrees of strained coherence (e.g., envy), preciousness (e.g., pardoning) or both. Self-accusation doesn't seem to come equipped with a corresponding role involving the other sides of justice as a relation between people, as if we had installed a judge who could *only* accuse and condemn and somehow lack the authority to exonerate or pardon.

But if the point of this comparison is to ground the bindingness of some normative demand upon us, then I think we have to take seriously the thought that we would have no reason to respect or feel bound by an actual judge whose authority was simply one-sided and implacable in this way. (Even God is traditionally taken to be more of a genuine Other than this, and therefore able to favor, reward, and redeem as well as to judge, torture, and damn.) What the one-dimensionality of conscience as the internalized Other suggests to me is that in trying to frame normative bindingness within the confines of a dialogue with oneself we are borrowing from those fragments of the structure of the normative, as it exists in our relations to existing Others, only those parts of it that will fit comfortably within the narrower range of relations to oneself that can be described without strain on what we might call the moral grammar of this family of concepts. And in doing so we are of necessity working with a truncated concept of the normative (such as the voice that can only accuse) in which we cannot recognize our actual commitments and obligations. Put briefly, for the work we want this concept to do, there is no substitute for an *actual* Other. In making these remarks I realize that I have only responded to part of the picture that Bagnoli is presenting, both here and in other writings, so I present these reservations, as they say, "in the spirit of dialogue".

Richard Moran is Brian D. Young Professor of Philosophy at Harvard University, USA, where he is currently Chair of the Philosophy Department. In addition to his book, *Authority and Estrangement*, he is the author of several articles in philosophy of mind, moral psychology, and aesthetics, in particular "The Expression of Feeling in Imagination", "Seeing and Believing: Metaphor, Image, and Force", "Anscombe on 'Practical Knowledge'", "Problems of Sincerity", and "Getting Told and Being Believed".

Address: Department of Philosophy, Emerson Hall, Harvard University, Cambridge, MA, 02138, USA. E-mail: moran@fas.harvard.edu.