Philosophy and Phenomenological Research
Vol. LXXXI No. 3, November 2010

Pre´cis of Self-Knowledge and Resentment[1]

akeel bilgrami
Columbia University

The term 'privileged access', on the lips and keyboards of philosophers, expresses an *intuition* that self-knowledge is unique among the knowledges human beings possess, unique in being somehow more direct and less prone to error than other kinds of knowledge such as, say, our knowledge of the physical world or of the mental states of others. These notions of 'directness ' and 'immunity to error' do, of course, need to be made more precise and may need more qualification (and even revision) than is provided at the level of intuition. Those are the familiar tasks of the philosophical *refinement* of an intuition. But these tasks must nest in a more basic philosophical question, which is to consider, as with all intuitions, whether the intuition can be *justified* in the first place by philosophical argument or whether, on scrutiny, it should be discarded as insupportable.

Self-Knowledge and Resentment addressed the intuition of privileged access in the limited domain of self-knowledge of intentional states, such as beliefs and desires. Its large conclusion, argued over five chapters, was that the intuition could be redeemed philosophically if we acknowledged in general, the close and integral relations between four different notions –value, agency, intentionality, and self-knowledge, and in particular the irreducibly normative nature both of human agency and of the intentional states of human agents. Without such an acknowledgement, it is more plausible and more honest to concede (to those who are skeptical of the soundness of the intuition) that self-knowledge is not distinct, except in matters of degree, from these other forms of knowledge.

The book begins with a characterization of two properties of intentional states which amount to the special character of self-knowledge.
1) *Transparency*, a property possessed by first order intentional states

[1] Akeel Bilgrami, *Self-Knowledge and Resentment*, Harvard University Press, 2006.

(restricting myself, as I said, to beliefs and desires) and 2) *Authority*, a property possessed by second order beliefs about the existence of these first order intentional states. Beliefs and desires are transparent if –not as a matter of contingency, but by their very nature—they can be said to be known by their possessors. And a second order belief about the presence of a first order belief or desire is authoritative if, by its very nature, it can be said to be a true belief.

The first chapter of the book spells out how and why these two properties capture and make explicit such intuitions as we have about 'directness' and 'immunity to error' and seeks thereby to have substantially revised and qualified these intuitions. Authority, if true, would certainly capture something of what is intuited in the idea of 'immunity to error'; and the intuition of 'directness', which presumably has to do with the fact that paradigmatic cases of self-knowledge of intentional states do not require of their possessors that they undertake analogues to 'looking' or 'seeing' or 'checking' as ordinary *perceptual* knowledge of the world does, is captured by the idea that it is because of their very nature rather than via these cognitive activities, that intentional states are known (to their possessors) –something we would not say about physical objects and facts, nor even about intentional states as they are known by those who do not possess them.

If these properties do make explicit the special character of self-knowledge, a bold initial move would be to begin by putting down two conditionals in a stark form, one for each property:

(T): If one desires or believes that p, one believes that one desires or believes that p.

(A): If one believes that one desires or believes that p, then one desires or believes that p.

These primitive conditionals are then accounted for in four chapters – (T) in chapters 2 and 3, and (A) in chapters 4 and 5, and in the accounting each is qualified in various ways that I will explain below.

The book's argument by which it is established that these properties hold of intentional states such as beliefs and desires turns on a preliminary point of central importance.

There is a deep ambiguity in the very idea of intentionality. It is widely (though not universally) thought that beliefs and desires are states that are in some sense deeply caught up with *normativity*. But they are also widely thought to be *dispositions* to behaviour. As some –for instance Saul Kripke[2]– have pointed out these are not entirely

[2] Saul Kripke, Wittgenstein on Rules and Private Language, Harvard University Press, 1982.

compatible ways of thinking of them. Much needs to be sorted out about this and the book does so at length (especially in Chapter 5). What emerges from it is the need to disambiguate the terms 'belief' and 'desire', making clear whether we mean to be talking of normative or dispositional states, when we use these terms. So, for instance, the term 'desire', when it describes an urge or a tendency I have, might be understood to have the dispositional sense. But it need not always be used to describe my urges and tendencies. It may be used to describe something more normative, something I think that I should do or ought to do. This latter is 'desire' qua *commitment*, not disposition. Thus an intentional state of mind that we might describe as the 'desire that I smoke a cigarette' could be an urge or a commitment on someone's part, and there is a distinction of principle between intentional states, conceived as one or other of these. (The desire that I smoke can, of course, be both an urge –or tendency– *and* a commitment, but in being so it is two things, not one, and that is why the term 'desire' is genuinely ambiguous.) As with desire, so with beliefs. Beliefs can be viewed as dispositions, which when they nest with desires (also conceived as dispositions) *tend*, under suitable circumstances, to cause behavior describable as appropriate to the propositional contents by which those beliefs and desires are specified. But they can also be viewed as commitments. Thus, a belief that there is a table in front of me is a commitment I have. If I believe it, I *ought* to believe various other things that are implied by it, such as, for instance, that there is something in front of me, or (more materially) that if I run hard into it, I will be injured. It is a commitment in the sense that it commits me to believing these other things, even if I don't actually believe them, just as my desires commit me to do various things, even if I don't do them.

How does a commitment contrast with dispositions (our urges and tendencies)? To put it in very brief summary: A commitment, being a normative state, is the sort of thing we can fail to live up to, even frequently fail to live up to, without it ceasing to be a commitment. After all it is in the nature of norms that we might fail to live up to them. By contrast, the very existence of a disposition would be put into doubt, if one did not act on it, if what it was disposed or tended to bring about did not occur (given, of course, the suitable conditions for its occurrence). When we fail to live up to a commitment, even under suitable conditions for the performance by which we live up to it, it does not put into doubt that one has the commitment –rather, all that is required is that we try and do better by way of living up to it (quite possibly by cultivating the dispositions necessary to live up to it.)

This disambiguation of the very notion of intentional states is important not only in itself but because the properties of transparency

and authority distribute quite differently depending on whether intentional states are conceived as dispositions or as commitments. How so? Authority holds of second order beliefs *only* if the first order inten- tional states they are about are conceived as commitments. Transpar- ency holds of first order intentional states, whether they are conceived as commitments or dispositions, but it only holds of the latter under a crucial further condition –it holds of those dispositions that are tied to one's agency, where by 'agency' I mean a notion of accountable human action, itself conceived in thoroughly normative terms. Thus, under this condition, transparency has wider scope of application since it takes in a wider class of mental states.

In the book, transparency is considered first.

That intentional states, conceived as commitments, should be transparent is due to the very nature of commitments. I had characterized commitments above, as requiring that we try and do better to live up to them, when we fail to do so. That, in part, is what makes a commitment, a commitment. If that is so, then I cannot fail to know my own commitments since I cannot try and live up to something I do not know I possess. But transparency, as I said, holds not just of intentional states conceived as commitments –it also holds of dispositions. And it is here that the relevance of notions of human agency and responsibility enters. The relevance is elaborated in the book by a modification and application of the innovative ideas in P. F. Strawson's essay ''Freedom and Resentment''.[3]

Strawson had argued that human freedom and agency are not non-normative metaphysical ideas having merely to do with issues of causality. Rather they are constituted by the *normative* practices surrounding notions of responsibility, such as blame and punishment, and these practices are, in turn, grounded in our normative reactions ('reactive attitudes' such as resentment and indignation) to each other's behaviour.

I extended this line of thought on freedom along the following lines to the notion of self-knowledge. For Strawson, the freedom of human action is a *presupposition* of our practices surrounding responsibility and the reactive attitudes that underlie them. To blame or resent another is intelligible only to the extent that he or she is capable of free action, and the blame and resentment only targets those free actions. To blame or resent a particular action is to presuppose that it has been freely enacted. My extension of this insight is this: Free and accountable human action, in this Strawsonian sense, *in turn, presupposes* that each such action is *also self-known.* And if that is so, the intentional states (whether conceived as commitments or dispositions) that

[3] See P.F. Strawson, Freedom and Resentment and Other Essays, Methuen 1974.

potentially go into the production of such action, are also self-known. In short, any intentional state of mind of a human agent that is tied (or potentially tied) to her actions which are the (potential) targets of justified reactive attitudes, is necessarily known to that human agent. To put it differently, we cannot justify having reactive attitudes (say, resentment) to actions / intentional states that are not self-known. This last may seem controversial since it seems to rule out any moral-psychological counterpart to the legal idea of strict liability, but a range of considerations are presented in the book to justify taking such a view.

Transparency, argued for along these lines, holds of intentional states qua dispositions. I am justified in resenting intentional actions (for instance those that cause harm) that flow from someone's dispositions, only if she has self-knowledge of those dispositions. Thus intentional action flowing from dispositions (that is, flowing not just from one's commitments but also from one's urges and tendencies) is free and accountable in Strawson's sense, so long as the dispositions are self-known. If we are justified, say, in blaming and resenting certain actions, then those actions (if Strawson is right) are free, and (if I am right) are self-known as are the intentional states (even if conceived as dispositions) from which they flow.

Transparency can now be fully characterized in the following refinement of conditional (T): To the extent that an intentional state is part of a rationalization (or potential rationalization[4]) of an action

[4] In this context, I use the term 'potential' here and elsewhere in the text, to talk about intentional states rationalizing actions, and it is a very general term that can cover a lot of things. But it should be obvious that by 'potential' in this context I mean something very specific and tightly controlled. By an intentional state 'potentially' rationalizing an action, I mean an 'intentional state, if *in its present status* in the moral psychology of an agent, were to rationalize an action, which it has not actually so far done.' What I do not mean by it is, 'if it were to rationalize an action which it has not actually done so far, *after having altered its status*.' I mention this for the following reason. Mental states which are not self-known have a status *different* from the states whose potential to rationalize I am claiming is caught up with agency. Yet these unself-known mental behavior states may *come to be* self-known by cognitive (e.g., psychoanalytical inquiry) and then they too might rationalize, which they have not actually so far done. When they do become self-known and when they then actually rationalize an action, those actions would be the object of justifiable reactive attitudes. So while they are still unself-known, in one sense of the term they still have the 'potential' to rationalize actions that are the objects of justifiable reactive attitudes. However, they would have this potential only in the sense that in order for the potential to be actualized, they would have to first change their status from unself-known to self-known, otherwise the actions they rationalize would not be the objects of justifiable reactive attitudes. That is a sense of potential quite different from the one I intend. What I intend is a distinction between actual and potential within the same status of intentional states. Perhaps one should drop the word 'potential' and find another, if this distinction is easily lost sight of.

or conclusion, which is or can be the object of justifiable reactive attitudes, or to the extent that an intentional state itself is or can be the object of justifiable reactive attitudes, then that intentional state is known to its possessor. Since the antecedent 'to the extent that…' relies on considerations of agency (as deriving from the Strawsonian ideas I mentioned), we can abbreviate the conditional for the sake of convenience and apply it to beliefs and desires in particular, as follows: *Given agency*, if someone desires (believes) that p, then she believes that she desires (believes) that p. This conditional (T) captures our intuitive idea that *by their very nature*, intentional states are self-known to their possessors.

I repeat: in this conditional, intentional states such as beliefs and desires may be conceived as commitments, but, *with the crucial antecedent in place*, they can be conceived as dispositions as well. Thus the proviso about agency, understood along Strawsonian lines, in the antecedent, is essential to this more capacious scope of transparency.

Authority next. Authority, the idea that our second order beliefs about our first order intentional states are always true, has seemed to many philosophers to be a very tall claim, given the widespread fact of self-deception and other Freudian phenomena.

The book seeks to provide an argument for why we may concede the ubiquitous fact of self-deception and other such phenomena (a concession that distinguishes 'authority' and 'privileged access', as I and others who have written recently about self-knowledge present it, from traditional Cartesian claims) while denying that that fact undermines authority. Here is a necessarily brief and rough version of the argument.

When one believes that one believes (or desires) that p, and one is self-deceived, it is not that one *lacks* the first order belief (or desire) that p, and therefore it is *not* that the second-order belief is false, it is rather that one *has another* first order belief (or desire) which is *not consistent* with the belief or (desire) that p (let's say, taking the clearest case, not-p). And, if the second order belief is not false, then this strat- egy has provided a way of viewing self-deception such that it leaves authority intact.[5]

[5] My claim here cannot be faulted on the grounds that it attributes *blatantly* inconsistent intentional states to an agent (just in order to save an agent's authority), and that it therefore is a violation of the principle of charity which forbids one to attribute blatantly inconsistent attributes to an agent. Blatant inconsistencies fall afoul of charity because there are no explanations given of why the inconsistencies exist. But when there are explanations for why there is an inconsistency, there is nothing uncharitable about attributing it. Sometimes the explanation is that the subject is unaware of one of the inconsistent beliefs. At other times, a subject may be (severally)

To put it less abstractly, let's take a standard sort of case of self-deception. Suppose someone has the following second order belief: she believes that she believes that her health is fine. But let's suppose that her behaviour suggests to others around her that she is full of anxiety about her health. Let's suppose that she does not recognize her behaviour as being anxious in these ways, but any analyst or even friend can tell that it is so.[6] One view to take of this sort of familiar case is that her second-order belief is simply false. It is the simpler view, and it is wrong. I think the right view is more complicated: her second order belief is true, which means she has the first order belief that her health is fine, but she also has another belief that she is not aware of, the belief that she is sick (or might be sick). So authority is not unsettled by the phenomenon of self-deception, rather *transparency* is missing regarding one of the two inconsistent beliefs (i.e., it is missing of the belief that not-p; in our example it is missing of the belief that she is sick). Perhaps she has suppressed her belief that she is sick because it is discomfiting to her to think of herself as sick, or because she does not want to be bothered with it in her busy life, and so on. If this is right, then allowing for self-deception clearly does not undermine *authority*. And, at the same time, for reasons mentioned in footnote 4 below, we have saved it from being undermined without any lack of charity in the attribution of inconsistent beliefs to the agent, since lack of charity in inconsistent attribution only holds if (among other things) the person is aware of both inconsistent beliefs. If she is unaware of one of them, it cannot be uncharitable to be attributed an inconsistency. In the example above, we have even offered specific possible explanations for the lack

---

aware of two inconsistent beliefs but has not brought them together, having compartmentalized them and their surrounding implications. And so on. In the case of self-deception, there will always be some such explanations of the inconsistency invoked by my strategy for saving authority. In cases of self-deception, it is perhaps most often (though not necessarily always) the former explanation that is in play. Assuming it is in play, we can admit that though it would be uncharitable to say of someone that she has inconsistent beliefs if she has self-knowledge of both the beliefs involved, in the inconsistency, in our example both of the inconsistent pair of beliefs are *not* self-known. In particular, the belief that not-p, mentioned above, is not self-known to the agent. It is not a belief, transparent to its possessor. And if that is so, there is *no* lack of charity involved in attributing inconsistent beliefs in this way to save authority since lack of charity only holds of cases of blatant inconsistencies, where there are no extenuating explanations of them in terms of lack of transparency of one of the inconsistent beliefs, or in some other terms.

[6] Often such a person may have a *half*-awareness of her anxiety regarding her health. Though in the book, I do discuss *grades* of self-knowledge while discussing self-deception, for the sake of brevity and simplicity, I won't here discuss cases of half knowledge that someone might have in such cases of her belief that they she might not be healthy.

of awareness of one of the pair of beliefs in the inconsistency, so there is nothing uncharitable about finding her inconsistent.

Of course, there is an immediate and obvious point regarding this strategy. First a bit of terminology: call the first order belief (the belief that one is healthy) in our example an 'embedded' belief since it is what the second-order belief takes as its object. In my strategy, in order to save the *authority* conditional I have allowed that the *transparency* conditional does *not* hold for the first order intentional states which are inconsistent with the 'embedded' first order intentional states of the second-order beliefs, whose authority is saved. This strategy saves authority of second-order beliefs about first order intentional states by insisting that in cases of self-deception these 'embedded' first order beliefs are indeed always present and therefore the second-order belief is always true –it's just that in each such case there is always another first order intentional state which is inconsistent with the 'embedded' first-order state and which is *not transparent* to its possessor. But to admit such a lack of transparency is all right since I have said that transparency (as captured in (T)) holds only when the proviso for agency, in the Strawsonian sense, holds –and we can grant that the relevant intentional states fail to meet that proviso.

But a question now arises, why should one deal with self-deception along the lines I am suggesting rather than as a less complicated phenomenon which is incompatible with the claim that we have first person authority over our intentional states? The answer lies in the intrinsically normative nature of intentional states conceived as commitments, as I have characterized them earlier.

As I said at the outset, authority holds only of first-order intentional states conceived as commitments and not dispositions. If we keep faith with the distinction between commitments and dispositions, we can say this: the behavioral evidence that is evidence of self-deception does not provide any evidence that the person lacks the *commitment* which is the 'embedded' intentional state of his second-order belief. It only shows that he has not lived up to the commitment in his behavior. His behavior reflects some of his dispositions, which of course he may not be aware of. And these will conflict with his commitments. All we need to find in order to attribute the *commitment* to him, is that when and if he does become aware of his dispositions and notices his failures to live up to his commitments, he accepts criticism for not living up to his commitments, and tries to do better by way of living up to them, by perhaps cultivating the dispositions to do what it takes to live up to them, etc. And so, even when he is not aware of his dispositions and his failures, so long as he is *prepared* to accept criticism etc. were he to become aware, that is sufficient to attribute the commitment to him. If

he meets these conditions for having the commitment, (i.e., if he has this preparedness), his behavior can no longer be seen as evidence for his second order belief being false, only of him not having lived up to his commitment.

Why exactly does the behavior not *also* refute the claim that he has the first-order belief, *qua commitment*, that he is healthy, thereby falsifying his second-order belief that he has such a commitment? Here is another way of putting my argument that makes it more explicit why not. Let's stay with our example and add that the protagonist not merely has the second order belief that he believes that he is healthy, but that he *says* he believes that he is healthy, i.e., he *avows* the first-order belief. (There is an elementary distinction between second-order beliefs and avowals that should not be lost sight of –the latter are not second-order beliefs, they are expressions of second order beliefs in words.) Now, two things must be established to conclude that there is authority: his avowal must be sincere (otherwise there is nothing –there is no second-order belief– to be authoritative since avowals express second order beliefs only if they are sincere avowals) and he must have the first order belief being sincerely avowed, which, of course in turn, requires that the defining conditions for his having the first-order commitment, must be met. Let us assume that the avowal is sincere, despite the behavioral evidence, because if it were not, there would be no question, as I said, of something being either authoritatively true or being false, and hence there would be nothing to dispute since authority is a property of second-order beliefs. Assuming the avowal to be sincere, we must ask, what are the conditions that would establish this sincerity of his avowal, given the behavioural evidence which suggest anxiety on his part about his health? The answer here is crucial and highly revealing: there can be no conditions which would establish the sincerity of his avowal *which would not also be the conditions which establish that he has the commitment he is avowing*. The conditions for having the commitment, I had said earlier, would be his preparedness (were he to become aware that he is not living up to his commitment) to accept criticism for not having lived up to it and his preparedness to try and do better by way of living up to it. These preparednesses, I am now saying, are the *very conditions* which would establish that his avowal of the commitment is sincere. What else could establish its sincerity?

So, if a sincere avowal is an indication that one has a second order belief that one possesses an intentional state, then it follows that our second order beliefs are always true because the conditions which allow us to say that she has the second order belief (that her avowal is sincere) are the very conditions under which we say that she has the

intentional state she avows. To the extent that it has been established that an avowal of an intentional state is sincere and, therefore, that a second-order belief really exists, then (even in the cases of self-deception), so must the intentional state it is about really exist, thus making the second-order belief true. No doubt, an agent may make insincere avowals. But what that shows is that we don't really have a second order belief since sincere avowals and second order beliefs stand or fall together. And if there are no second-order beliefs, then the subject of authority is not yet on the table, since authority is a claim about the truth of second-order beliefs, not the truth of insincere avowals. But, if and when authority *is* on the table, self-deception need not be seen as overturning it. Second-order beliefs need not be seen as having any role in a psychological economy without the presence of the first-order beliefs they are about.

On this basis, I concluded that (A), the conditional for authority, is established, but its reach is more limited than (T) since, on the argument I have offered, it holds only of first-order intentional states, conceived as commitments, not dispositions.

That summarizes the refinements the book made on the intuitions regarding privileged access –showing the intuitions to be captured in two properties of intentional states such as beliefs and desires that are, in turn, captured in two conditionals, and giving arguments for the truth of those conditionals.

As I said, the argument only goes through for the property of authority, *if we assume that intentional states are themselves normative states such as commitments*, and though the argument for transparency goes through for both commitments and dispositions, it only goes through for the latter, *if we assume a normative notion of agency that owes to Strawson's notion of freedom and modifies it in one fundamental aspect*. Those are both large assumptions on large topics, and since they each drive the two arguments for the special character of self-knowledge via these two properties of authority and transparency, I will close this pre´cis of the book, with a very brief indication of why I claimed we should make both those assumptions.

1) For the first assumption, it is important to understand the idea that normativity is central to intentionality in a particular way, in a way that has it that intentional states such as beliefs and desires are *themselves normative states* (since that is what the idea of commitments are). Davidson who was something of a pioneer in arguing for centrality of normativity to intentional states (and thereby repudiating various forms of naturalism about intentional states, such as physicalism and functionalism) fails to see just this point and despite his claims for the relevance of normativity to intentionality, he views beliefs and desires as

dispositions, not commitments. For him the normativity allows these states to be dispositions but views these dispositions as being 'governed' by normative principles (principles of deductive, inductive, and decision-theoretic rationality). That, by my lights, is insufficient and a good part of Chapter 5 presents reasons for why we need something stronger by way of normativity, viewing beliefs and desires not as first order dispositions governed by normative principles, but rather commitments that are themselves normative states. To establish this, an argument is needed against the naturalistic equation of intentional states with dispositions. The book offers what I call a 'pincer' argument for this stronger (than Davidson's) claim, which in (far too brief summary) is this.

One arm of the pincer invokes and adapts G. E. Moore's open question argument that targets the reduction of value or norms to natural properties in general, to a more specific target: the reduction of intentional states to dispositions in particular, which are, as Kripke rightly points out, states that cannot be thought of normatively and can only be given a naturalistically descriptive characterization since they are causal *tendencies*. The relevance of the open question to a view which takes beliefs and desires to be dispositions would be roughly that someone can always *non-trivially* ask: ''I have all these dispositions to /, but *ought* I to /?'' If this is a genuinely non-trivial question, if it is not like asking, say, ''Here is a bachelor, but is he unmarried?'', then that would suggest that intentional states such as beliefs and desires are internal oughts (commitments) not to be reduced to first order dispositions.[7]

The other arm of the pincer is motivated by a limitation of the first arm. The Moorean argument works only if one assumes that there is a *definitional* equation of intentional states with dispositions. But much of contemporary philosophy of mind has aspired to something much less strong. It has been quite satisfied with something like an assertion of an a posteriori identity of intentional states with dispositions, on the model of other a posteriori identities such as water=H_2O or Hesperus=Phosphorus. Here the Moorean argument will not be effective since it targets only definitional reductions. These identities, being a posteriori, turn not on the meaning or definition or 'sense' of the terms involved ('water', 'Hesperus', etc) but on their reference, usually

[7] I say first order dispositions deliberately. Second-order dispositions may well be involved in the characterization of commitments. In characterizing commitment, I say that failures to live up to commitments require of an agent that she tries to do better by way of living up to them; and it might well be asked if this requirement is satisfied by the exercise of a disposition to try and do better. I can allow such sec- ond-order dispositions, pointing out that it does not amount in any way to reducing

elaborated in the last forty years or more in causal-theoretic terms. So, the second arm of the pincer drops the Moorean considerations and invokes at this stage a Fregean argument to supplement it. The argument has a familiar pattern. Someone can deny that intentional states are dispositions (or, better, deny that some particular intentional state is some particular set of dispositions –even when it is identical with it) without being inconsistent or irrational. But if that is so, then to account for the fact that such a person's mind represents a completely consistent state of affairs, the terms on each side of the equation being denied will need to have a sense over and above a reference. If one restricts oneself to the reference or extensions of the terms, the person would seem to be inconsistent. But we know that he is not. He merely lacks some information, he fails to know a worldly identity. So just as the terms 'water' and 'Hesperus' would need to have a sense if we were to make it come out that someone who denied that water=H_2O or denied that Hesperus=Phosphorous was not being irrational and inconsistent, we will need to posit that the intentional term in the identity or equation being denied by him has a sense. But this raises the question: what is the sense of the intentional term expressing?

Here we have a choice in answering this question, a choice that amounts to a dilemma for the naturalist who wants to equate the intentional state with a naturalistic property like a disposition. Either it is expressing a naturalistic property or it is expressing a non-naturalistic property. If it is expressing the latter, then of course, it straightforwardly undermines naturalism. So one assumes that the naturalist will insist on the other option and claim that it is expressing a (further) naturalistic property. At this stage, the first arm of the pincer re-asserts its relevance and closes in on the naturalist once again. For now, if it is the *sense (or meaning or definition)* that is given in terms of the naturalistic property, then that is precisely what the Moorean open question consideration is once again effective against. Moore's argument, as we said, is geared to target definitional reductions.

Thus a Moorean argument, supplemented by a Fregean argument, together construct a pincer effect against the naturalistic equation of intentional states with dispositions. We start with Moore, then introduce Frege to deal with a posteriori identities, which in turn *returns* us to the Moorean argument, if the naturalist appeals to senses that express *naturalistic* properties. And the effect of such a pincer argument is to make room for the assumption that my argument for authority requires, viz., that intentional states are internal 'oughts' or commitments, not dispositions.

2) The assumption of a normative notion of agency, which presupposes that one's intentional states (whether conceived as commitments

or dispositions) are transparent so long as they fall within the purview of such agency, owes to Strawson's re-orientation of the notion of agency towards a norm-based metaphysics. Without it, the presupposition of transparency would not go through.

Strawson was speaking to a traditional debate about human freedom in which two opposing doctrines shared a common background commitment –that freedom was incompatible with the universal sway of causality. Determinism (or 'hard determinism' as it was sometimes called), one of the two opposed doctrines took the view that universal causality put into doubt that freedom was so much as possible, while Libertarianism, the other doctrine, took the view that the fact of freedom depended on a 'contra-causal' capacity of the human subject or will which put into doubt that causality did have universal sway.

Strawson rejected the shared background commitment of these two opposing doctrines and thereby formulated a version of what is often described as 'compatibilism'. But his compatibilism was quite different from traditional forms of compatibilism in introducing an explicitly normative element that they lacked. Traditional ways of resisting the shared background commitment took the form of saying that though causality may be universal, not all causes were coercive or 'compulsive' or 'constraining' causes (to use Hume's terms). Those which were coercive causes thwarted human freedom, but many causes were not coercive and that left open the possibility of free human action even within universal causality. One can understand Strawson's version of the doctrine of compatibilism as emerging out of a criticism of this more simple version of it. Suppose we ask the question: what about a coercive cause makes it coercive and what about a non-coercive cause makes it non-coercive? His view would be that just staring at the causes in question won't help to answer this question. We have to look at our practices of such things as blame and punishment, and their underlying moral-psychological basis, which consists in our reactive attitudes of resentment, indignation, etc., to even so much as identify which causes of actions are coercive and which non-coercive. It is not as if causality (i.e., the distinction between coercive and non-coercive causes) is irrelevant to freedom, it is rather that there is no identifying these causes as distinct types of causes without appeal to some *normative or evaluative* considerations such as our practices of blame and punishment and the reactive attitudes that underlie them. Thus for instance a harmful act that issues from a non-coercive cause would go hand in hand with our attitudes of, say, resentment towards the act, whereas an act that issues from a coercive cause goes in tandem with our attitudes of excusing that act. It is not as if one identifies the coerciveness and non-coerciveness of the causes of the act

independently of these attitudes towards the act, and then comes to have these attitudes on the basis of that identification. Rather these occur together. There is no norm-independent identification of these causes as distinct types of causes.

This innovative move on Strawson's part was a real advance in the philosophical account of human agency, but I had argued (in Chapter 2) that it stops a little short of the full extent of the normative dimension of agency that is needed. The uncompromisingly committed determinist might still argue that what Strawson presents as the deepest grounds of human agency –our reactive attitudes– are themselves unjustified. For such a determinist, given the fact of the universal sway of causality, our moral psychology in which the reactive attitudes figure so centrally is indulgently judgmental, and determinism requires that we should really be suspending our reactive attitudes. This would, of course, in turn affect how we conceive of the practice of punishment (since for Strawson that is grounded in the reactive attitudes), which would now be thought of on a more medical model, something purely instrumental, a model of 'repairing someone' rather than blaming them, and reacting to them with attitudes of resentment and indignation.

Strawson's predominant response to such a view in his celebrated essay is to frankly and simply say that this is to fail to understand who we are. We cannot imagine a human life that is a life entirely rid of a moral psychology in which the reactive attitudes are central. In my discussion, I quote passages where Strawson makes this response and I argue that it is complacent on his part to simply plunk down the unimaginability of such a pervasively judgement-free mentality. People under conditions of alienation (whether from social or psychological sources) often don't care to be judgmentally reactive and we *can* imagine a comprehensive extension of such a condition that will exemplify the determinist's scenario of kicking the ladder of agency (of the reactive attitudes) away from under one. And even if we cannot perhaps easily *achieve* such a comprehensive surrender of agency, we *can* decide to commit such agential suicide *by* committing *biological* suicide. So long as the underlying motive is to commit the former, that still leaves it as a moral psychological possibility that we can actualize.

If suspending the reactive attitudes is not unimaginable, how, then, might we justify the possession and the retention of the reactive attitudes (and therefore, of our agency) against the extreme determinist who asks us to suspend them as far as we can? I argue that we can justify the reactive attitudes (and, therefore, agency) not by going to something *more* fundamental and general than agency, but from *within* agency itself. In other words, we need not try and justify the reactive

attitudes and the agency they ground foundationally. We can justify our being agents with reactive attitudes, that is agents who are normative subjects, by citing *particular* norms or values that they further. Thus specific norms and values we have can justify the much more general idea of the very possession of norms and their exercise in judgments and reactive attitudes. This is an internalism in justification, a form of normativist coherentism of practical reason to match coherentism of beliefs in theoretical reason, where propositions of a high generality may be confirmed by propositions whose content is specified in much specific terms.

How is this insistence on my part that we must go further in the normativist direction than Strawson's stopping point relevant to my account of self-knowledge, in particular my account of the property of transparency that intentional states (whether thought of as commitments of dispositions) possess? In other words, what role does my further demand for the *justification* of the reactive attitudes themselves play in accounting for transparency?

Strawson does not need nor want further justifications of the reactive attitudes because he merely claims that *freedom* is presupposed whenever the reactive attitudes are in play. But I want to say not merely that freedom is presupposed when the reactive attitudes are in play, *but self-knowledge (transparency)* of intentional states is also presupposed. Now, there is a common view that we may have reactive attitudes of resentment (and even blame and punishment) towards someone who does another harm unself-knowingly. One of the reviewers of my book says this: ''When the self-deceived person harms another out of spite, we find fault with more than her *ignorance-of- the-harm* she causes, but also fault her *spitefulness*.''[8] If this is right, then resentment and blame do not presuppose self-knowledge on the part of the subject who is resented and blamed. In Chapter 3, I try and demonstrate at length that though we do often *have* such reactive attitudes, there is no *justification* for such reactive attitudes, when we have them. But to even so much as raise this issue, we have to raise the prior issue as to whether and when the reactive attitudes themselves are justified. And to raise that issue is to be set on a path, a quite general path, that takes one further down the normativist path I described earlier, than anything found in Strawson, who shuns that path by saying we cannot imagine not doing without the reactive attitudes, so there is no question of seeking some justification of them.

[8] Krista Lawlor, 'Review of Akeel Bilgrami, *Self-Knowledge and Resentment*', Mind, vol. 117, April 2008.

These assumptions 1) and 2) that I have been elaborating of the deep and radically normative nature of both intentionality and of agency reach out by strict implication to other large themes and claims in the book, such as for instance, (in Chapter 4), the first person point of view, its distinctness from the third person point of view[9], the sceptical implications of that distinction for various forms of reduction of intentionality, including even to something as weak as supervenience…. Since I expound and defend these additional claims in my 'Replies' to Baldwin and Normore below, I won't spell them out here, but I will say this.

One cannot establish what makes self-knowledge unique among the knowledges we possess unless we see it as having these distant connections to wider themes in the philosophy of mind and the moral psychology of agency. It is the book's presiding claim that it is the network of relations that self-knowledge bears to these detailed and radically normative elements of agency and intentionality that allows one to account for self-knowledge without turning to any perceptual or other routine forms of epistemological explanations. Indeed, it goes further than other 'constitutive' accounts of self-knowledge by denying that even some of the recent talk of the 'entitlement' to self-knowledge on the basis of our first order intentional states giving us reasons for the relevant second order beliefs, has any particular aptness within the sort of account on offer here.[10] Such talk has real bite and point when one is pursuing a more substantial epistemological project, such as is found paradigmatically in a perceptual account. Philosophers who have discarded the perceptualist model should be discarding this kind of residual talk of 'entitlement' as well. In perceptual knowledge there is a crucial element of a *dynamic transition* involved in the warrant that is provided by facts and objects in the world for our veridical perceptual beliefs about them. I make a much more radical claim than other 'constitutive' views of self-knowledge precisely because I don't think it is apt to say that self-knowledge involves a dynamic transition in which our first order intentional states give us *reason* to form our beliefs about them. Though it is true that neither the concept of a reason, nor even the concept of an entitlement, *as such*, imply such a dynamic transition, the very specific 'reasons' claim (sometimes made by

[9] In the book, I use the expressions ''The first person point of view'', ''The agentive point of view'' ''The point of view of engagement'' synonymously, as also, by contrast, the expressions, ''The third Person Point of view'' and ''The detached point of view''.

[10] See particularly the contributions of Christopher Peacocke and Tyler Burge in the symposium ''Our Entitlement to Self-Knowledge'' in *Proceedings of the Aristotelian Society*, 1996.

philosophers) mentioned in the last sentence, which underlies the idea of an entitlement to self-knowledge on the basis of our possessing intentional states, *does* imply it. Claims of that specific sort and the rhetoric of 'entitlement' they have generated —reflecting, as they do, this dynamic element rather than stressing, as I do, not the dynamic but the *integral* connections that self-knowledge bears to a range of normative notions that characterize our agency and our intentionality— has no suitable place in my account.

But, then, this puts a *great* burden on what I have called the 'integration' of self-knowledge with these other notions, so much so, that it is only a slight exaggeration to say that the book, by its end, presents four problems, sometimes even called 'mysteries' by a certain kind of naturalist, that have vexed philosophers for so long – the possibility of agency and freedom in a deterministic universe, the place of value or norm in a world of nature, the relation between intentional states and the central nervous system, and the special character of self-knowledge—as really, in one sense, at bottom, the *same* mystery. At any rate they are so highly integrated that there is no understanding any one of them without coming to grips with all.

If one thought instead that self-knowledge, being *knowledge* after all, was just another narrow epistemological theme, I don't think we could account for our intuitions about privileged access. Viewed in purely epistemological terms, without integration with questions of agency, norm or value, and the irreducible nature of intentionality, the widespread cases where we manifestly lack self-knowledge of our intentional states (such as self-deception, for instance), would make these intuitions seem like outdated Cartesian dogma. There is something honest, then, about those who refuse to grant anything special to self-knowledge and view it as getting a causal account based on a measurably more than usual *reliable* mechanism that will account for our intuitions misleadingly expressed as 'privileged access'. They see it as a narrow question in epistemology, they find the exceptions to be ubiquitous, and they draw their conclusion that there is nothing radically set apart about self-knowledge. Their conclusion is honestly drawn from their frame- work. It is their framework that is wrong. Self-knowledge *is* unique *only if* it is embedded in a much wider framework integrating very large themes in philosophy that my book traverses. Why should we pursue it in a broader rather than a narrower framework? I will put the answer to this question flamboyantly: because it allows us to reduce four mysteries to one. In philosophy, surely that should count as some kind of progress.

Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
Vol. LXXXI No. 3, November 2010

Fool's Good and other Issues: Comments on *Self-Knowledge and Resentment*

calvin g. normore
McGill University

In ''Freud, Morality and Hermeneutics'', Richard Rorty drew an analogy between the way in which the mechanical philosophy of the seventeenth century had successfully reshaped our vocabulary and our ways of thinking and the way in which psychoanalysis promised (or threatened) to do so.[1] He advocated that we try out Freud's spiffy new approach which dispensed with concepts like blame and punishment in favour of concepts of therapy and adjustment. Nietzsche had argued a century earlier that much of our normative vocabulary was the product of mean-spirit and crabbiness and best dispensed with. Why not, one might think?

Akeel Bilgrami's *Self-Knowledge and Resentment* is a formidable response to Rorty's challenge. Beginning from Peter Strawson's claim in ''Freedom and Resentment'' that we cannot abandon reactive attitudes like resentment because they are central to what we are, Bilgrami works out an account of what it is to be so. Unlike Strawson, who regards a project like Rorty's as simply impossible, Bilgrami thinks it perfectly possible; just as we can commit biological suicide we could commit what he calls 'agential suicide' (p.60) and he is agnostic about whether we could have the agential suicide without the biological. He is adamant, however, that it would mean giving up normativity and, he argues, that, if carried far enough, it would entail giving up mentality itself

Self-Knowledge and Resentment is an intricate and sustained argument that there are items, minds and states of mind, for example,

[1] Rorty, Richard ''Freud, Morality and Hermeneutics'' in *New Literary History* Vol. 12, No. 1, Psychology and Literature: Some Contemporary Directions (Autumn, 1980), pp. 177–185.

which are real and irreducible and which constitutively involve a perspective on them—the perspective of the first person. At its heart is the thesis that most of our states of mind—of our thoughts and desires in particular - are *commitments* and that to have a mind is thus something which can only be characterized in irreducibly normative terms. Bilgrami also insists that a commitment requires a *preparedness* to accept criticism and to cultivate the dispositions which would lead to living up to the commitment.

What then is the connection between the normative and what can be characterized in non-normative terms? For Bilgrami such claims as that the normative supervenes globally on the non-normative are unassessable because they require us to adopt simultaneously two stances toward the world—that of an agent and that of a scientist / spectator. It is here that my central disagreement with Bilgrami lies. Where he sees two incommensurable stances I'm inclined to see just one—that of a participant *in* the world—and various abstractions from it.

Bilgrami criticizes John MacDowell for holding that the connection between intentional states and non-intentional effects can be understood as a causal connection. He maintains that this commits MacDowell to claiming a univocal notion of cause at work both in intentional and in non-intentional causal connections while, in fact, that involved in the non-intentional cases is deeply connected with that of covering laws while that involved in the intentional cases is not (p.247). I'm with McDowell. Hume may have thought that we would not be prepared to say that A caused B were it not that occurrences of events of the same sort as A were regularly followed by occurrences of events of the same sort as B but I don't find good reason to believe it. Once that is given up it is very hard to see exactly what laws have to do with causality. In fundamental Physics causal laws are very hard to find at all—the emphasis is rather on equations and symmetries - and in Philosophy most counterfactual theories of causation, for example, do not privilege laws. Moreover, I don't see how laws could help us understand causality nor how adding generality to an explanation makes it a better *explanation*. Perhaps there are some whose curiosity is assuaged when told that this crow is black because all crows are, but my own is only increased. Hard enough it is to figure out why this crow is black—considerably harder to figure out why they all are ! Bilgrami writes that ''The explanation that comes with citing causes, if it is indeed the same notion, does seem to bring with it some demand for treating like cases alike and not as distinct singularities. Generalities are built into the idea.'' (p. 261) But I doubt it to be so. Consider an example (due to Anscombe and to Feynman): I put a piece of radioactive material under your bed connected to a Geiger counter which in

turn is connected to a bomb. If we get enough clicks on the Geiger counter the bomb goes off. Suppose it does—then the setup just described caused your death. Yet as similar a setup as you like could be in place and no bomb go off. Once we give up on determinism, like causes need not have like effects. That does not mean the notion of cause has changed and if, indeed there is a notion of cause which is neutral between intentional and non-intentional cases, then *as agents* we can affect the non-intentional world.

Bilgrami is of the mind that there are two points of view, the first person (or agential) and the third person (that of the scientist / spectator), and that one cannot straddle them. As he sees it ''The spectator views the world in a detached way, including scientifically and predictively, and while he does so he cannot view it as making normative demands on him to act or even to ask what ought I do….The agent views the world in a deliberative, first person mode and asks how he ought to act, but while doing so he cannot view it in a detached, including scientific and predictive, mode'' (p. 254).

Bilgrami has us imagine a being who has a third person point of view but not a first person point of view. Such a being ''is simply blind to those facts in (and aspects of) the world that natural science does not study''. (p. 254) He does not have us imagine a being who has a first person point of view but not a third person point of view. Strawson's related exploration (in *Individuals)* suggests that a being lacking a third person point of view would also lack a first person point of view in the sense in which you and I have one—a perspective on things other than ourselves. That encourages me to think that the third person point of view is an abstraction from a richer one which includes what we access as agents. That in turn encourages me to think that the issue is not, at Bilgrami suggests, whether we could straddle two perspectives as self-contained as two Spinozist attributes would be, but rather whether we could at the same time occupy the whole of our conceptual space and only a part of it.

Engineering, Medicine, and plain helping out all normally require taking both an agential and a third person point of view. From your perspective you see me as having (say) needs in virtue of my low caloric intake. You can indeed see the caloric intake itself as calling on you—it's too low—less than required to keep me in good health - and you conclude something must be done about it.You plan to bring me meals on wheels. You are five kilometers away. You are 15 minutes away as you drive and ten minutes away as your partner drives. My low caloric intake is partly a product of circumstance and partly of a genetic disorder. There is nothing about the third person perspective that you need to leave out in any of this. Bilgrami points to Spinoza's

remark that one cannot deliberate while at the same time predicting how one will act (p. 251). This may well indicate that not every way of developing the third person perspective is compatible with the first person perspective, but it remains crucial to deliberation that we incorporate in it a third person point of view. ''What if I were to do this - then that would very likely happen and that and that. I'd better think again.''

In a related vein one might wonder whether the contrast between the first person and the third person is too stark. Thomas Nagel (in *The Possibility of Altruism*) brought to our attention whether, there might be features of the world detectable only by subjects but nonetheless intersubjective and public in the sense that different subjects might be similarly related to them and might recognize this. Nagel's example was your pain. There are no unfelt pains and if you are in pain this provides a reason for *you* to do something about it. Nagel argued that it also provided a reason for *me* to do something about it though I do not feel it. *Either* of us might fail or refuse to acknowledge that reason but it would be there for both of us and we could both recognize this.

Nagel was operating with a conception of reason at some distance from Bilgrami's. On Bilgrami's view as I understand it, reasons are typically commitments and normally one does not have a commitment without acknowledging it. It is that dynamic which underlies the two conditionals he takes to characterize self-knowledge. I'm less clear than he about this. I seem to have commitments (to family and others) I did not undertake and commitments I undertook but do not acknowledge. My commitments are transmitted along lines of transfer—by logical principles and others—that I do not fully understand. I take commitments not to be in general transparent to the agent who has them but am open to the thought that reasons are. If this turns out to be so then, while my commitments are transmitted by modus ponens whether I realize it or not, my reasons are only transmitted via modus ponens if I do realize it. Values seem to me to play both sides of that street. When I come to recognize the value of something I don't think it suddenly has a value it didn't have before but only when I *see* that value can it figure in my reasoning and be a reason for action or belief.

Pains do not appear in the vocabulary of Physics (though they do in that of current Neurophysiology) but to identify the third person perspective with that of Physics may itself be an oversimplification. Other sciences employ concepts which seem to straddle the first and third person perspectives. Affordances in J.J. Gibson's sense are identified in terms of a (human or non-human) animal's possibilities for action. Is it then from a first-personal or a third-personal perspective that one does Gibsonian animal psychology? If we suppose it is a

third-person perspective then there can be items very like values which we recognize from the third personal perspective. They don't call on us from that perspective—we have abstracted away from those aspects of them and of ourselves on which they might call—but we don't have to suppose on that account that an affordance is a different thing when we consider it as agents and as scientists. On the other hand, if the concept of an affordance only makes sense from an agential perspective then we may need that perspective to do much animal psychology. Again the distinction between the agent and the scientist would be under siege.

Bilgrami seems concerned that if we do not compartmentalize the agential and scientific views and so render unassessable such theses as that the normative supervenes globally on the non-normative, we will be driven to assert the primacy of a third person perspective. Given the recent history of Philosophy his concern is certainly reasonable. Still, I'm less concerned; even within the natural sciences intervention is as central as representation. In the light of Bilgrami's arguments one might well ask instead what it is exactly that prevents the integration of the natural sciences within a perspective that takes the first person seriously. Although there certainly are particular theories within the natural sciences that seem hard so to integrate it is not at all clear (at least to me) that there is anything in the methodology or basic structure of the natural sciences which would make such integration impossible.

Central to Bilgrami's argument for the claim that the agential and scientific points of view are distinct and incommensurable is what he sometimes refers to as a pincer strategy (p.218). One jaw of the pincer is what he calls the Moore-Kripke argument; the other is the thesis that certain terms, 'good' for example, must have Fregean senses. The goal of the strategy is to rebut naturalisms—efforts to make the normative and intentional derivative on what is accessible from the perspective of the spectator.

Bilgrami calls the one jaw the Moore-Kripke argument because he takes it that G.E. Moore's 'open-question' argument is at the core of Kripke's argument that rule-following cannot be reduced to acting on dispositions. The central idea, as Bilgrami understands it, is that since it is always a non-trivial question to ask ''I have this disposition to / but ought I to / ?'' the mental state involved in thinking I ought to / is not the disposition itself (and neither is the rule). Bilgrami is clear that the Moore-Kripke argument will only work directly against definitional reductions and he is sensitive to the charge that if it worked generally we could not have a posteriori identifications like ''heat is mean molecular motion'' but he thinks that invoking this just drives one

against the other jaw of his pincer—the argument that evaluative terms must have Fregean senses. Here he argues:

''But if someone can coherently and meaningfully say ''It is not the case that good is x'', then we must ask what the term 'good' in this coherent, meaningful false statement means. Let the term *denote* whatever it is supposed to, given the a posteriori identity; the point is that we need to posit a sense, we need to say what it *connotes*, in order to find the statement coherent and rational, even if it is false. That shows that even in the identity statement ''good=x'', the term 'good' has a sense, over and above a reference. (p. 217)

I agree with Bilgrami that there are good reasons to think the normative and the intentional not reducible to what is not normative and not intentional but I confess to being skeptical of the strength of either jaw of the pincer. On the one side Kripke's argument against the identification of rule following with acting on a disposition does not seem to be of a piece with Moore's argument. Whereas Moore is concerned to show that because it is non-trivial to ask for any term X non-synonymous with 'good' ''This is X but is it good?'' the property of being good cannot be identified with any other property, Kripke is concerned to show that to settle whether one is acting on a given disposition does not determine *which* rule one is following. While Kripke's argument, if successful, establishes that a rule is not simply a disposition, it does not establish that following a rule could not *inter alia* essentially involve acting on a disposition or being in some other non-normative state. On the other side of the pincer it is not at all clear to me why we need a *sense* for 'good' to find the statement ''It is not the case that good is x'' coherent. After all we do not need a sense for 'Cicero' (or for 'Octavian'' to find the statement ''Cicero is not Octavian'' coherent? When pressed on this Bilgrami is inclined to stress the special character of intentional states and to suggest a more intimate connection between them and the way they are accessed than is the case for most objects. (p. 374 fn 12) To bolster this he imagines a strict causal theorist simply refusing to admit that there are senses involved in our use of 'good'. Such a position entails, he suggests, the epistemic possibility that we be systematically mistaken about what is good. He continues: ''But about good, it seems utterly unacceptable to think that it is possible that we have never got it right. *It is unacceptable in the sense that in the face of such a consequence, we would be perfectly within our rights to say that the interesting normative notion is 'fool's good' and not good.*'' (p.220)

This seems to me too quick. It is indeed difficult to imagine that we could be systematically mistaken in identifying a familiar property. Still, systematic error about goodness seems easier to imagine than, say, systematic error about heat—precisely because there is much less

agreement about goodness in the first place. In one tradition it is good to kill one's innocent son if God commands it, in another such killing is paradigmatically evil. Even with heat systematic error seems possible -—although the chili pepper certainly felt hot it turned out to be no hotter than the rice used to assuage the feeling. To suppose that we have a special certainty about either the property of goodness (if there be one) or about its extension is, I make bold to say, what Bilgrami would call complacence.[2]

None of this touches Bilgrami's most central claim– that self-knowledge and first person authority are grounded on the constitutive connections among those of our commitments which have as their objects our beliefs, desires and intentions. Bilgrami's contribution here is of the first importance. In its light I'm ready to embrace, believe and be committed to his conclusion that we are, if not *au fond* at least very deep down, normative beings. I only wish I understood better what it is so to be.

[2] Though even if complacence it be we should not be complacent about it. cf. Gampel, Eric ''Ethics, Reference and Natural Kinds'' *Philosophical Papers* vol. 26 (1997) no. 2 pp. 147–163.

Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
Vol. LXXXI No. 3, November 2010

Comments on A. K. Bilgrami's *Self-Knowledge and Resentment*

thomas baldwin
University of York

Bilgrami begins his book about self-knowledge with a self-deprecating warning which suggests that the book is going to be a hard slog through some tough analytical reasonings without much that is likely to be of interest to 'non-philosophers'. This is very misleading: the book is engaging and accessible and the book's conclusions are of great interest. Bilgrami argues that we will not be able to understand ourselves properly unless we appreciate that self-knowledge is integral to our capacity for free agency and that this capacity cannot be subsumed within the understanding of the world provided by the natural sciences, even though there are other aspects of ourselves which are to be understood in this way. So, as Bilgrami recognises, he is endorsing a dualist account of the self: in one respect we are rational agents whose actions involve normative commitments which we cannot make without knowing what we think; in another respect we are animals affected by natural causes whose effects on us may well not be known to us. Dualism is a notoriously unpopular position these days, but Bilgrami attempts to show that his dualism of 'perspectives' (1st person vs 3rd person, as he calls it) is not vulnerable to the familiar objections levelled against dualisms of the past. As this review will indicate, I am not persuaded that this is so; but the challenge provided by Bilgrami's stimulating book certainly takes the discussion of this issue in new directions.

Bilgrami does not present his new dualism up front at the start of the book; it only emerges in 'full bloom' (p. 266) towards the end as he unifies and extends themes from the preceding chapters. Much of the book is directed to presenting and arguing for what he calls a

'constitutive' conception of self-knowledge, in particular of knowledge of one's own beliefs, desires and intentions. This conception of self-knowledge is contrasted with perceptual and inferential accounts which conceive of self-knowledge as comparable to our knowledge of the physical world; Bilgrami argues that accounts of this kind imply 'a certain *independence* of the first-order mental states from the sec- ond-order states about them', and it is precisely this independence which he rejects in affirming his constitutive conception of self-knowl- edge. One can see what Bilgrami has in mind here, though it is odd that he says nothing about 'response-dependent' conceptions of sec- ondary qualities, which indicate that even within a perception-based account of knowledge of the physical world there is room for some- thing comparable to the constitutive thesis he wants to advance; the contrast between perception and constitution is not as sharp as he presents it as being. What is more pressing, however, is the need to get clear just what his constitutive account amounts to. In part, of course, it is just a denial of the independence of beliefs, desires and intentions from second-order beliefs about them, and Bilgrami makes this explicit by setting out two conditions which frame the central chapters of the book: the first, 'transparency', affirms that if someone believes or desires that *p* then he believes that he has this belief or desire; the second, 'authority', affirms that if someone believes that he believes or desires that *p*, then he believes or desires that *p*. These two conditions together imply that first-order states and beliefs about them are mutually interdependent; this point, however, cannot be all that a 'constitutive' thesis amounts to, since constitution is an asym- metric relation – it cannot be that *x* constitutes *y* and that *y* consti- tutes *x* (unless different kinds of constitution are involved). So, we need to ask, which of the conditions (transparency, authority) is the basic constituting one? Bilgrami is not as clear on this matter as one would like, but as the book progresses it becomes clear that it is the authority condition that is fundamental. Bilgrami takes it that the paradigmatic phenomenon is the sincere avowal of beliefs, desires and intentions ('I believe ..', 'I want ..', 'I intend ..'). The avowal expresses the speaker's second-order belief but equally commits the speaker to having the mental state in question, and in this way 'constitutes' it. I am not sure how much Bilgrami intends his readers to read into his use of the term 'authority', but there are many situations in which the exercise of authority is constitutive, and Bilgrami seems to treat sincere avowal as comparable to an Austinian performative, where

the speaker's exercise of a constitutive authority is certainly essential.[1] Given this approach, it is now easy to see how self-knowledge comes 'for free', as Bilgrami puts it: since avowal constitutes the mental state avowed, the second-order belief expressed by the avowal is both true and appropriately warranted; so the speaker knows that he has the state that he avows.

What then of the transparency condition, that one know that one has the beliefs (etc.) that one has? Plainly, this condition is satisfied for whatever beliefs one has constituted by an authoritative avowal of them, but one would commit the old fallacy of affirming the consequent if one took it that this was a reason for asserting the transparency of belief. Bilgrami is of course innocent of this fallacy; instead he argues that transparency is a pre-requisite of responsible agency. The argument is familiar enough: the familiar *mens rea* requirements of responsibility imply that the agent understood what his intentions and beliefs were in acting as he did. So, as Bilgrami puts it, 'given agency', the transparency condition follows. Bilgrami takes it that this way of thinking about transparency shows that the self-knowledge involved is of the constitutive type and cannot be just the outcome of a perceptual / causal relationship between the agent and his thoughts. For, he holds, relationships of this latter kind are liable to break down, but their breakdown is not sensitive to the question as to whether or not the agent was responsible for what he did (p. 122). This argument is not persuasive. A sensible causal theorist will hold that the causal relationship which underpins self-knowledge is inherent in agency itself. The 'reasons are causes' thesis familiar from Davidson, Pears and many others implies that there is a causal component in rational agency, and all that the theorist who maintains that self-knowledge is fundamentally a causal matter has to hold

[1] At this point I want to register a complaint on behalf of Gilbert Ryle. In chapter 1 of his book Bilgrami criticizes Ryle for failing to do justice to the special status of self-knowledge on the grounds that he (Ryle) holds that self-knowledge is just a matter of coming to understand oneself in much the way that others come to understand one. This is of course an important strand of Ryle's position (as it is of Bilgrami's since he takes the same view about knowledge of one's own natural dispositions). But it was Ryle who also introduced the conception of avowals as a different way of thinking about self-knowledge (see *The Concept of Mind* p. 183), and since Bilgrami makes this concept the keystone of his position he should have given Ryle some credit for it. In other respects too, Ryle's account of self-knowledge is much richer than Bilgrami suggests: for example, Bilgrami might with profit have used Ryle's Humean thesis of the 'elusiveness' of the self to enrich his discussion of 'transparency'.

is that this causal component in rational agency includes the causal relationship between the reason (belief / desire) and the agent's second-order belief which is required to secure transparency. So far from being an *ad hoc* manoeuvre to avoid Bilgrami's objection, this refinement strikes me as a sensible way of developing a causal account of agency which accommodates his point that responsible agency requires transparency. Indeed the result is a causal theory which, so far from having a problem concerning transparency, provides the requisite self-knowledge 'for free'.

Bilgrami's account of transparency is preceded by a discussion and elaboration of themes from Strawson's famous paper 'Freedom and Resentment'. The connection between self-knowledge and resentment to which Bilgrami's own title adverts to is not obvious at first, but it turns out to be central to the dualism that emerges towards the end of the book. Bilgrami holds that there is an fundamental connection between agency and self-knowledge: as we have just seen, the transparency condition rests on the thesis that agency is sufficient for self-knowledge, and, as we shall see below, Bilgrami takes it that the authority condition rests on the thesis that agency is also necessary for self-knowledge. The thesis that Bilgrami advances in this chapter is that agency is inherently normative: it is tied to reactive attitudes such as resentment which have an irreducible reference to values. Hence once this thesis is added to the 'agential' conception of self-knowledge, it will follow that self-knowledge is inherently normative too; and Bilgrami's dualism concerning the self is then inferred from the dualism of the value / nature distinction.

Plainly there is much here that is disputable and I shall question some of Bilgrami's later claims below. But the thesis that agency is a capacity whose characterisation involves irreducible reference to value-judgments is uncontentious. Nonetheless, in presenting and elab- orating it, Bilgrami makes some questionable claims and it is worth briefly examining these. According to Bilgrami, it is in Strawson's paper that this normative aspect of agency is first clearly identified, and Bilgrami contrasts Strawson with Hume's famous 'reconciling' discussion 'Of Liberty and Necessity' on this point. But once one turns to Hume one finds that Hume emphasizes exactly the same point: 'For as actions are objects of our moral sentiment, so far only as they are indications of the internal character, passions, and affec- tions; it is impossible that they can give rise either to praise or blame, where they proceed not from these principles, but are derived

altogether from external violence.'[2] Thus Hume, like Strawson, takes it that it is our moral concern with the 'internal character' of persons which leads us to differentiate unfree from free action. For Hume (and, I believe, Strawson) this point is compatible with a causal understanding of agency; but here Bilgrami would disagree. For he holds that once one accepts that free agency is a normative concept, one must deny that it is a matter of one's action being caused in one way rather than another. It is certainly right that one cannot provide a conceptual account of agency exclusively in causal terms; but it does not follow that there is not a causal ingredient in a conceptual analysis, nor that, as a matter of fact, the distinction between free and unfree action is a distinction between different causes of action (Bilgrami would contest this final point by means of an argument to which I return later).

In addition to this misrepresentation of the relationship between Hume and Strawson, there is a further mistake in the way in which Bilgrami presents himself as extending Strawson's position by emphasising the way in which agency is of value to us. Strawson, he suggests, was right to use the values inherent in our reactive attitudes to ground the distinction between free and unfree action, but then 'stops at the fact that we have reactive evaluative attitudes' (p. 67) when what is needed is a further recognition of the ways in which agency is of value to us – which Bilgrami illustrates by alluding to the varied valuable activities of the members of the Bloomsbury Group (p. 63). Yet Strawson in fact says: 'the personal reactive attitudes rest on, and reflect, an expectation of, and demand for, the manifestation of a certain degree of goodwill or regard on the part of other human beings towards ourselves.'[3] So Strawson does take the matter further by connecting the reactive attitudes to the moral 'demand' inherent in our relationships with others. This is a type of 'interpersonal' consideration which Bilgrami largely neglects; the value he associates with agency is

---

[2] *Enquiry concerning Human Understanding* ed. T. Beauchamp (Oxford: Oxford University Press 1999) pp. 161-2. In this connection it is worth adding that Bilgrami's comment that Hume was 'perhaps' the first philosopher to advance a compatibilist strategy for dealing with the issue of free will is very wide of the mark. Much the same strategy had been proposed by Hobbes in his letter (published in 1646) to the Marquis of Newcastle with the title 'Of Liberty and Necessity' (it cannot be an accident that Hume uses the same title); and both Hume and Hobbes advert to earlier discussions in scholastic philosophy of the way in which human freedom and divine necessity can be reconciled. Thus what is interesting here is the way in which early modern philosophers adapt the positions developed in earlier theological debates to debates within 'natural philosophy'.

[3] 'Freedom and Resentment' p. 14 in *Freedom and Resentment and other essays* (London: Methuen, 1974).

essentially first-person, - the value for each of us of realising our own projects. For Strawson, by contrast, the reactive attitudes structure our moral relationships with others, and it is this interpersonal domain that is fundamental to the characterisation of free agency.

I mentioned earlier that Bilgrami's thesis of the constitutive role of self-knowledge is based on the phenomenon of avowals. He acknowledges (pp. 158-9) that a causal theorist might seek to accommodate this phenomenon, and therefore sets himself to show that once first-order belies, intentions and desires are properly understood, we will be able to understand fully how it is that our constitutive authority with respect to these states detaches them from a purely causal understanding of their role. Bilgrami's discussion of this point is complicated and wide-ranging, and I am not confident that I have fully mastered it; but, as I read him, the key *gestalt* switch he seeks to induce is one from thinking of these states in a functionalist manner in which they are defined by their causal roles to seeing them as theoretical and practical *commitments* around which we structure our deliberations concerning what to think and do. A key element in Bilgrami's discussion of this point is his insistence that deliberation, and rational thought generally, is an activity, and thus that agency is necessary for thought of this kind. Animals are of course capable of synthesising and using perceptual information to advance their goals, but by and large their behaviour does not have the intentional structure which would make reactive attitudes to it appropriate, and it is of course agency of this kind that Bilgrami has in mind when affirming that agency is necessary for thought. The conclusion he then draws is one in which the three themes of normativity, agency and self-knowledge are 'integrated' as he likes to put it: the first-order thoughts of a rational agent are normative commitments made by the agent who makes them precisely by his second-order avowal of them. So, as he puts it, 'the conditions for any *particular* second-order belief are the very same as the conditions for the particular intentional state it is about. If, therefore, the second-order belief exists, so must the first-order intentional state it is about. In other words, the authority conditional is true' (p. 159).

If I am right the key element of this position is what Bilgrami calls the 'commitmental' (p. 304) conception of first-order intentional states, - beliefs, intentions, and desires. On the one hand, a causal theory really does look to be inadequate to account for commitments, and, on the other, it makes good sense of our supposed constitutive authority with respect to these states to think them as commitments. I myself think that this is the right view to take of beliefs and intentions, and the phenomenon which demonstrates this point is Moore's paradox. The incoherence of the thoughts *I believe that it's raining but it isn't*

and *I intend to go to New York but I'm not going to do so* reflects the fact that the belief and the intention I attribute to myself bring with them commitments, to the fact that it is raining and to my going to New York, commitments which I then explicitly repudiate.[4] But it is then notable that there is no similar variant of Moore's paradox which applies to desires: there is nothing incoherent in the thought *I want to have another cigarette but I'm not going to do so*; and this, to me, suggests that Bilgrami is mistaken in thinking that his 'commitmental' conception of mental states applies straightforwardly to desires. Of course we often express our desires in the first person way by saying 'I want …' but this does not show that we thereby have any constitutive authority with respect to them. What is more significant is the role of desires in practical deliberation, for insofar as we treat them as providing reasons for action this does seem to imply that we treat them as having some normative significance. But I think we restrict this role to desires which we endorse: someone seeking to stop smoking can acknowledge the existence of a desire for a cigarette without taking it that this desire provides him with a reason for acting on it. If this is right, then it is not desires *per se* which have normative significance for us, but those which we endorse. Since endorsement is a commitment to the positive value of that which is desired, this conclusion provides a route back to something approaching Bilgrami's position.

Finally I turn back to Bilgrami's dualism. In the arguments I have reviewed he connects the themes of intentionality as commitment, self-knowledge, and agency, and combines his discussion of these themes with a critical approach to causal accounts of them. In his long final chapter he attempts to provide decisive arguments for this contrast between scientific, causal, accounts of these phenomena and his own value-laden approach to them by introducing anti-naturalist arguments from ethics. Not surprisingly he draws especially on Moore's 'open question' argument, which he takes to imply that putative causal-scientific conceptual reductions of phenomena such as agency and intentionality are bound to fail. This is relatively uncontentious but in a bold move to which I alluded before he seeks to extend Moore's argument to undermine causal-scientific accounts which do not pretend to provide a conceptual analysis but which nonetheless seek to identify the underlying relevant conditions in their own terms. Bilgrami argues that accounts of this kind fail because they fail to show how the

[4] I argued for this thesis in 'The Normative Character of Belief' in *Moore's Paradox*, eds. M. Green & J. N. Williams (Oxford: Oxford University Press, 2007) pp. 76-89. The insight that Moore's paradox applies to intentions as well as to beliefs is due to Andre´Gombay: see his neglected masterpiece 'Some Paradoxes of Counterprivacy', *Philosophy* 63 (1988) 191-210.

disputed property, such as goodness in Moore's own case, can be conceived both as a natural property and as a property of some conceptually different type. For, Bilgrami asks (p. 218), 'What if we say that the relevant sense <sc. whereby goodness is conceived> does not express a naturalistic property? Well, then, there is no fully effective naturalistic reduction in the first place.' This argument is unconvincing: there is a crucial ambiguity in the phrase 'does not express a naturalistic property' between 'does not express a property in naturalistic terms' and 'does not express what is in fact a naturalistic property'. The conclusion follows only if the latter interpretation, that the property is not in fact naturalistic, is assumed. But the starting point is that given by the first interpretation, according to which goodness is supposed to be being conceived as a moral property which, as Moore's open question argument shows, is not a way of conceiving goodness in naturalistic terms. Since there is no way of inferring the second interpretation from the first, the argument fails. Conceiving of goodness as a moral property is not conceiving it as a natural property; but it does not follow that it is not a natural property. As a result some of Bilgrami's anti-naturalist claims seem to me unsubstantiated, and I am even less persuaded by his rejection of Moore's thesis that, despite not being a natural property itself, goodness supervenes upon natural properties. For what is striking about Bilgrami's rejection of supervenience is the way in which the point depends on his radical dualism of perspectives. According to Bilgrami, we each have available to us a first-person perspective in which we express our evaluative commitments, our reactive attitudes and exhibit the constitutive authority of our self-knowledge; equally, we can adopt a detached third-person perspective within which none of our commitments or attitudes are expressed, and in which we view the world and ourselves from a neutral point of view which includes, but is not exhausted by, a scientific understanding of things. The thesis of the supervenience of goodness (or whatever) on natural properties therefore implies that there is a dependence relation between properties which belong to these different perspectives: for values belong to the first-person perspective, natural properties to the third-person one. And it is this implication which Bilgrami asserts to be incoherent, simply on the ground that these perspectives cannot be combined.

This last claim is astonishing. One would normally think that one can introduce impersonal, more-or-less scientific, facts into a first-person evaluative perspective; indeed it seems a truism of practical deliberation that we do this all the time when we consider the best means to achieve our ends. Bilgrami attempts to refute this objection by arguing that as soon as we bring what we might think of as

detached scientific considerations into our practical deliberations, we are bound to reinterpret them in a value-laden way which is appropriate to their role within a first-person perspective (pp. 256-7). This is completely unpersuasive: suppose mathematical calculations are involved in my practical deliberations – does it then follow that the mathematical truths involved have to be reinterpreted as value-laden first-person truths? This is absurd. Incidentally, it is worth observing that it is no part of Kripke's anti-naturalism, which Bilgrami endorses, to hold that this is so. For Kripke, there an irreducible normative dimension to mathematical concepts, but it does not follow that he is committed to accepting that there is a similar normativity inherent in mathematical truths (and if one were at all tempted to affirm this paradoxical thesis, it would follow that all truths are in this way normative, and thus that there could not be the detached, third-person perspective which Bilgami rightly upholds). I suspect that what drives Bilgrami to his extreme position is the fear that if he allows scientific and other similar natural facts to be incorporated as such into a first-person practical perspective, he will have to allow the supervenience of values on natural properties and then he will find himself on a slippery slope to the reduction of values to natural properties after all. As I indicated above, I am not persuaded that Bilgrami has a good objection to non-conceptual reductions of this kind, but setting that aside it is certainly not settled that supervenience implies a reduction of any kind. Moore denied this, and I myself agree with him: supervenience is a consistency requirement, not a reductive thesis.

Bilgrami attempts to head off objections to his perspectival dualism by arguing that the forceful objections to dualism are objections to Cartesian mind / body dualism to which he is not vulnerable. Of course he is right to differentiate his position from that of Descartes (though he does affirm that if one is an agent, then one experiences oneself as such and thereby knows infallibly that one is an agent – pp. 197-8). But there are nonetheless severe problems inherent in the implication of his position that third-person scientific facts about my body and my mind cannot be incorporated into my first person perspective of myself as an embodied agent. Only a little experience of physical and mental illness or disability is needed to show that this perspectival separation involves a quite unintelligible alienation. In fact there is a different philosopher whose dualist position anticipates that of Bilgrami in many respects; Bilgrami alludes occasionally to Spinoza, but it is contrary to Spinoza's monism to hold that our ways of understanding ourselves both as an extended thing and as a stream of ideas cannot be combined. Instead the philosopher whose work I was continuously reminded of by Bilgrami's book is the J-P. Sartre of *Being and Nothingness*.

Bilgrami's first-person perspective captures Sartrean *being-for-itself* while his third-person perspective deals with Sartrean *being-in-itself*. This is not the place to show in detail how Bilgrami is vulnerable to the aporiai of Sartrean dualism, but one salient similarity concerns the way in which neither of them does justice to the interpersonal perspective of our normative relationships with others. Sartre notoriously finds that others can be grasped only as potential threats; Bilgrami's position is not as extreme as this – he can describe others and their values from within his third-person perspective. But what he cannot do is capture the point emphasized by Strawson, that our reactive attitudes rest on demands for mutual goodwill that we make of each other. For this is not a value inherent just in our first-person perspective: it is a demand made on us by others on whom we make a similar demand. It is belongs neither to a first-person nor to a third-person perspective, but to an interpersonal perspective which can incorporate elements of both the others alongside its own distinctive features.

Bilgrami's perspectival dualism is a dead end. But, as I have indicated, I share his 'commitmental' conception of beliefs and the constitutive conception of self-knowledge which he elucidates. So what is required is a new, post-Humean, reconciling project of showing how to combine these essentially first-person phenomena with both a full acknowledgment of the kind of third-person understanding provided by the natural sciences and an appreciation of the significance of phenomena such as our reactive attitudes which rest on the interpersonal demands we make of each other. This is a tall order! Throughout the last parts of his book Bilgrami develops his dualism by means of a critical debate with John McDowell and McDowell's writings certainly provide an important resource for fulfilling this project. For myself, however, I find that the writings of J-P. Sartre's old antagonist Maurice Merleau-Ponty (especially his *Phenomenology of Perception*) an even richer resource.[5] But I shall not attempt here to explain how this is so, and instead it remains for me to thank Bilgrami for the inventiveness and integrity of his wonderful book. As I have indicated, I disagree with many of his claims; but disagreement is the lifeblood of philosophy, and in his capacity to provoke creative disagreement Bilgrami shows himself to be one of the leading philosophers of our time.

[5] Several of the essays in *Reading Merleau-Ponty* ed. T. Baldwin (London: Routledge, 2007) address this theme.

Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
Vol. LXXXI No. 3, November 2010

Replies to Tom Baldwin and Calvin Normore

akeel bilgrami
Columbia University

I am grateful to Tom Baldwin and Calvin Normore for the trouble they have taken to write their fine commentaries on *Self-Knowledge and Resentment*. I will respond to their interpretative points as well as challenges in these ''Replies'' and am glad of the chance to try and make clearer and better some of the claims and arguments in that book.

Reply to Tom Baldwin

1. Baldwin begins by characterizing my position as a 'dualism' that claims: ''in one respect we are rational agents whose actions involve normative commitments which we cannot make without knowing what we think; in another respect we are animals affected by natural causes whose effects on us may well not be known to us.'' He, then, adds: ''Dualism is a notoriously unpopular position these days, but Bilgrami attempts to show that his dualism of 'perspectives' (1st person versus 3rd person as he calls it) is not vulnerable to the familiar objections leveled against dualisms of the past. As this review will indicate, I am not persuaded that this is so…''

Every point that is made by him in the quoted characterization of the 'dualism' is so innocuous (even, I would have thought, by his own lights) that one might wonder how it could possibly be that such a position is objectionable. But, as he says, he is not persuaded that it is not objectionable. So there must be some slippage between what he characterizes here and what he really views my position to be. I believe that there is such a slippage and I will return to it later –-since Baldwin doesn't return to discuss the 'dualism' till the end of his comment, I too will leave it till later; in fact, in order to avoid repetition, I will leave it to my next 'Reply' to Calvin Normore (who focuses mostly on

the 'dualism'), and address both Baldwin and Normore jointly on that theme there.

2. Baldwin then characterizes my view of what sets self-knowledge apart from other kinds of knowledge by expounding the contrast I make between it and our perceptual knowledge of the world. Self-knowledge lacks, he quotes me as saying, ''a certain independence of the first-order mental states from the second-order states about them'', an independence that does exist between the facts and objects in the world and out perceptual beliefs about them. I had (borrowing from Crispin Wright's terminology) used the word 'constitutive' to describe this particular lack of independence peculiar to self-knowledge. He adds that it is ''odd that he [Bilgrami] says nothing about 'response-dependent' conceptions of secondary qualities, which indicate that even within a perception-based account of knowledge of the physical world, there is room for something comparable to the constitutive thesis he wants to advance.'' But in my book I *do* discuss the response-dependent conceptions of secondary qualities. I actually even begin my discussion by pointing out precisely what he thinks I should have pointed out. I say: ''The judgments that are supposed to determine rather than track an independent domain of color facts are perceptual judgments, but it is my whole point (and Wright's too) that my second-order judgments about my beliefs and desires are precisely not perceptual judgments… '' (p.295).

I see, then, that I need to get across quite generally, but particularly to Baldwin, the thrust of my views that he has managed to skip here because they bear on his efforts at interpretative constructions on my behalf a little later in his commentary. Though I am grateful to him for those efforts, they get me quite wrong and one can see why that is so only if one is attentive (in the way he has not been) to the points I make about the disanalogy with Wright's idea of 'response-dependence'.

I was keen to say two things about the various phenomena (secondary qualities, intentional states, values, …) that are gathered by Wright as falling under the notion of 'response-dependence'. First that the gathered phenomena don't amount to some sort of *kind* that deserves a more or less uniform treatment. And, more important, second: if, as seems to be true on Wright's view, the point of introducing the notion of 'response-dependence' is to try and provide grounds for a very refined version of the doctrine of anti-realism, then, in the case of intentional states (and their self-knowledge), the notion of response-dependence does not apply very well at all.

At first sight, my denial of self-knowledge as something to be modeled on perceptual knowledge may seem to make matters better, not

worse, for the claim to anti-realism, since something which is not a matter of being perceived is more likely to be susceptible to anti-realist treatment. In general, the truth of a perceptual belief is taken to be *prima facie* at least, indicative of *facts* or *objects* that were perceived. So, it might be thought that, though it is true that what are gathered together by Wright as response-dependent phenomena are not uniformly treatable as a kind, what I have denied about intentional states (that they are paradigmatically perceived by their possessors in self-knowledge of them) makes *them particularly* apt for a response-dependent treatment that has as its motivation to establish an anti-realism about the relevant phenomenon. But that thought is spoilt by a complication. *Others can* know my intentional states *by perception* even if I don't know them by perception. And that presumably then *restores* a realism about intentional states. What, then, if someone protested by saying: ''Nothing is spoilt, others cannot really *perceive* my intentional states, they can only *infer* them from my behavior which they perceive.'' The protest seems to be granting that were others to be able to perceive my intentional states, a realism about them could be restored, but since they are not really perceiving them, only inferring them from what they perceive – a person's behavior– realism is not restored at all, and on the contrary an anti-realism is re-asserted. But now it seems that realism about some phenomena lies in the non-inferential availability of the phenomena to one's judgment.[i] If so, then one may after all restore the realism about intentional states once again by pointing out that the availability of my intentional states to *my* judgment is indeed non-inferential because of the constitutive thesis for which I have been arguing.

These considerations –there are others too which I won't rehearse here–put into doubt that it is sensible to see the constitutive view of self-knowledge of intentional states that I propose as promoting any- thing like the anti-realism generating ideal (however refined)) of response-dependence.

3. This has bearing on what Baldwin goes on to say immediately after in his comments. In an interpretative move, intended to be helpful, he says that my constitutive view cannot merely be (what I present it as being) —that self-knowledge is set apart from perceptual knowledge because of the *mutual inter*dependence that is established by my two conditionals, a mutual interdependence of the first-order intentional states and the second-order intentional states that

[i] This is a familiar criterion for anti-realism that Michael Dummett has presented in many writings of his. See my ''Meaning, Holism, and Use'' in, ed. E. Lepore, (Blackwell, 1986) for a discussion of this criterion and of Dummett's anti-realism more generally.

self-ascribe those first-order intentional states, a mutual interdependence that is missing in perceptual knowledge between the facts or objects being perceived and the perceptions of them. Baldwin thinks that the term 'constitutive' suggests something less mutual in dependence. What does the constituting must be primary in some way and so he proposes that I must think of the second-order beliefs that are authoritative as primary since they must be constituting the first-order intentional states that are self-known. He proposes that, of my two conditionals, I must see the one for authority, therefore, as more primary than the one for transparency, and claims even to detect hints of this primacy in my book. I don't see things this way at all and I am sorry if my use of the word 'constitutive' suggests any such primacy for authority or any such special power in the second-order beliefs that go into self-knowledge. I did not mean 'constitutive' to bring with it the metaphysical baggage of something doing some actual constituting, being, therefore, in that sense, primary. I use the term really only to distinguish the view from the perceptual paradigm of knowledge in which what is known has a certain form of independence from the knowing. That independence does not exist in self-knowledge because of a *mutual* interdependence of what is known and the states that carry the knowing. Each conditional brings out a different dependence in one of two directions and neither is more primary than the other.

I think Baldwin slips into this proposal on my behalf because he does not notice that I had distanced myself from the supposed affinity of my view of self-knowledge of intentional states with 'response-dependence' views of secondary qualities. Those views –their very name suggests it, as does the aspiration to an anti-realism– claim a primacy for the response (in the case of self-knowledge, for the second-order intentional state). If you do claim that, one can see why Baldwin should think that I must have in mind to make authority more central (since it is a property of second-order states) than transparency, thus introducing an asymmetry in my two conditionals, which I had presented as symmetrical. However, I don't claim that and so I refuse his interpretation of my view, even as I express thanks for having cautioned me to the misleading nature of the rhetoric of 'constitutive' that I had adopted to describe my view. I had thought that I had been quite stipulative in the way I had wielded that term, marking only the distance of my view from the perceptual paradigm (a distance owing to the mutual interdependence I mentioned above). But I can see that if one stipulates meanings for terms that in other contexts of usage seem to suggest something stronger than what has been stipulated, one can mislead a reader. The fault is partly mine then, though I insist that

Baldwin would not have been misled if he had paid attention to my repudiation of the assimilation of my view with response-dependence conceptions.

One last point on this theme. Baldwin suggests that I really, at bottom, make avowals absolutely central to the overall argument of my book. I do not. I mention avowals very briefly in giving one, among other, arguments to establish the conditional for authority. The Strawsonian argument for transparency is just as vital to my overall argument. Even within considerations of authority, avowals are not central. What are central are the second-order beliefs that are authoritative. Avowals, as expressions of second-order beliefs, only enter in a specific argument for authority because they cannot be sincere without the second-order beliefs they express being true beliefs since the evidence for their sincerity also establishes the presence of the first-order states that would make the second-order beliefs true. He then goes on to use this primacy he attaches to avowals to elaborate and justify my own claim that 'self-knowledge comes for free'. This is a mistake. I use that expression to describe *transparency*, not authority. Transparency is a property of first-order states. My point was that if a first-order belief or desire needs no effort of cognition to be known –not perceiving, not inferring, not even checking (as one might with the position of one's limbs)—then to know it requires nothing from us. In my view, just the fact of first-order states figuring in our *agentive* lives makes them known to us. Possessing them in the course of such lives *is* to know them. The slogan 'self-knowledge come for free' merely marks this fact –that it takes nothing from us to know them.

4. This discussion is followed by a misunderstanding in the way that Baldwin presents my opposition to what I call 'causal-perceptual' accounts of self-knowledge. I introduced the idea of such causal *accounts* of self-knowledge in a very specific way and, in opposing them, I was not opposing the idea that *causality* is involved in self-knowledge. The idea of such accounts was introduced as follows. Perceptual accounts of self-knowledge of intentional states, I had said, are essentially causal accounts since the 'perception' involved in *self*-knowledge does not involve 'looking' or 'seeing' or other such cognitive activities. 'Turning your eyeballs inwards' is merely a grotesque metaphor. The heart of the idea of self-knowledge by observation cannot be captured in such metaphors. It is rather captured by causal mechanisms. So, perceptual accounts of self-knowledge are more spare than perceptual accounts of our knowledge of the external world. Unlike the latter, they posit no cognitive activity over and above the underlying causal mechanisms relating what is perceived and the

perception, i.e., the relevant first and second-order mental states –and, of course, they need an overlay of justificatory epistemology. In other words, the perceptual account of self-knowledge is based on two things: a) a first-order belief or desire *causes* a second-order belief about it (about its presence or existence) and b) a justificatory element that elevates this second-order belief into knowledge, a justificatory element that is essentially reliabilist, i.e., the causal mechanisms involved are deemed to be highly reliable. I called this the 'causal-perceptual' account of the transparency property of first-order intentional states and I set about opposing such an account and replacing it with my alternative Strawsonian account.

In the latter, self-knowledge is a presupposition of agency and agency itself is a fallout of a larger picture in which justifiable reactive attitudes are central. I had raised the question whether this disallows any causal element and said it certainly did not, but unlike as in the causal element that underlies the perceptual account (as just expounded above), the *work of accounting* for self-knowledge was not done by the causal element (by a mechanism and its reliability), it was done by considerations of agency. Baldwin accurately presents a point I made about why the explanatory work was not done by the causal relations between the first-order intentional states and the second- order beliefs about them, when he says the *breakdowns* that are possible in causal mechanisms are breakdowns that could not possibly be expected to match the *absence* of agency since the latter is a matter of normative considerations of when reactive attitudes can be justifiably applied and the former is a matter of the failure of a mechanism. Failures of mechanism are simply not sensitive to the question as to whether or not the agent was responsible for what he did, if the latter is understood as Strawson suggests it should be. And it cannot be replied that the sensitivity is the other way round, i.e., the reactive attitudes are sensitive to the presence or absence of the causal factors. My Strawsonian view does not allow that we can establish the causal factors that are appropriate for the exercise of the reactive attitudes (resentment, say) that ground responsibility, *independently* of the normative considerations that lie in the reactive attitudes themselves. (There is more on this a little further below.)

Hence, on my view, it makes no odds to say that causality is built into agency, as Baldwin, invoking Davidson's authority, does. If agency is thoroughly normatively conceived and is doing the work of accounting for self-knowledge, then causal elements can be as present as you like (and I do like), they can be as *built-into* agency as you like, but

they will not be doing the explanatory work.[ii] The crucial point, therefore, is that the kind of co-variation between appropriate first- and second-order states of mind that is fixed once the proviso for agency in the conditional for transparency has been fulfilled, is *simply not hostage to causality*. I repeat, saying that it is not *hostage* to causality is not to say that it cannot be causal at all. But it does rule out the kind of causal *account* that underlies perceptual accounts of self-knowledge. That is why I call it, perhaps misleadingly as I admitted earlier, a 'constitutive' account.

5. On the prior question of freedom, quite independent of issues of self-knowledge, Baldwin faults my reading of Hume, which had said that Hume is insufficiently Strawsonian in the formulation of compatibilism. He does so by citing a passage from Hume that is supposed to place him in greater proximity to Strawson's innovations on the subject. But to think that that passage does that is to either misunderstand Hume or to misunderstand Strawson or some combination of these. Just as bad, what he proceeds to make of this in commenting on my own understanding of the relation between Strawsonian notions of freedom and the role of causality, is very wide of the mark.

The passage he cites from Hume can only have one possible interpretation in the context of what precedes it in the Humean text. Hume, in previous pages, has just traversed his reasons for saying that liberty is compatible with necessity. He defines liberty as issuing from 'a power of acting or not acting, according to the determinations of the will'. This, he has said, is not to be contrasted with necessity, that is, with the fact that our will is subject to our passions, our dispositions, our internal character, our affections…, it is only to be contrasted with notions of constraint (what I, in my 'Pre´cis' above, called coercive or compulsive causes) and he cites as an example the most extreme form of such constraint –someone being, like a prisoner, ''in chains''.

[ii] In the book I say that in a world in which there are no causal relations between first-order intentional states and second-order beliefs about them, and in which agency is understood as I claim it must be understood (along Strawsonian lines), there will be self-knowledge. Baldwin's point about seeing causality as being *built into* agency might be seen to be denying that such a world is possible. I find it hard to understand such a view of being 'built in' –the causal and the normative being yoked together *analytically* in some way. (Davidson, in a sphere of philosophy quite different from the present theme of self-knowledge, sometimes has suggested that the normative and the mental *dispositional* are analytically yoked together by his principle of charity, which disallows mental dispositions to come apart from ratio- nality. I discuss that argument in Chapter 5 of the book and try and show how limited its appeal should be.) Be that as it may, my point really can be made with- out even denying this idea of the casual element being built into agency, since even if it is built in, it is not doing the work of accounting for self-knowledge as it is in the perceptual account.

(Philosophers since Hume have stressed a spectrum of examples of constraining causes of less extremity, such as threats backed by violence that coerce people into doing things, and going so far, as Nozick had done, of considering as an example, bribing someone to do something when he desperately needs the money that the bribe offers.[iii]) Immediately before the passage cited by Baldwin, Hume, then, says that liberty, so defined, is a necessary condition for ('essential' for) moral blame. And in the cited passage itself, Hume adds that our blame for such actions requires that those actions flow from causes that are non-'constraining' (even if 'necessitating'), states of mind such as those I cited above: moral dispositions, passions, internal character, etc.

Now, there is nothing whatever in the passage that stresses what Strawson does. Strawson does not merely say that questions of liberty are *related* in this way to questions of blame and moral sentiments towards human actors. In fact hardly anyone denies that, not even most of those who think that freedom is a purely metaphysical issue (by 'purely' I have in mind something more traditionally austere than the *norm-based* metaphysics of freedom that Strawson had proposed). It is quite possible for an anti-Strawsonian of this sort to take up questions of reactive attitudes and blame and remain anti-Strawsonian, so long as she says that a necessary condition for our reactive attitudes and blame is a *purely metaphysical condition* (the condition that the blameworthy subject is acting on non-constraining causes such as the dispositions and passions and affections that Hume mentions as influencing the will to act.) The innovative thing that Strawson adds to this compatibilist picture is that there is no identifying *as a prior*, whether a condition that is appropriately necessary for a blamable act has obtained, without seeing the blame and the underlying reactive attitudes of resentment, etc. *itself* as part of what will rightly identify it as such. For Strawson, what makes a non-constraining cause a non-constraining cause and a constraining cause a constraining cause (internal character, dispositions, passions and affections as opposed to chains, mortal threats, bribes offering one money when one is desperate for money) is *not worn on the sleeves* (is not visible in these bare descriptions) of this laundry list of opposing examples. The examples are only properly identified as one or the other sort of cause by also bringing in our reactive attitudes, our practices of blame or of excusing, respectively. These can't be left out of the identifications. Cause and reactive attitudes are not merely related to one another. There is no identifying

[iii] Robert Nozick, ''Coercion.'' In Philosophy, Science, and Method: Essays in Honor of Ernest Nagel. Edited by Sidney Morgenbesser, Patrick Suppes, and Morton White. (St. Martin's Press, 1969)

the former as a prior, without the normative dimension of the latter being intrinsic to our understanding of which kind of cause it is. Hume says absolutely nothing of *that* sort in the passage Baldwin cites. He, therefore, says nothing of a distinctively Strawsonian sort.

In my book, I say very explicitly (page 54 is the location of only one such explicit statement) that this innovation of Strawson might give the impression that he is simply *reversing* the direction of explanation, i.e., Hume in that passage has it that we need to be clear that the right sort of cause (internal character, dispositions, passions, affections…) is present in order for us to praise or blame or make the actions 'the objects of our moral sentiment' and Strawson reverses things by saying that we should look first at the blame and the moral sentiments and derive in secondary fashion, the right sort of cause. I go on to say that that impression would be wrong. They are to be conceived *in tandem*, as conceptually linked in a single conception in which neither is prior and it is important that neither side of this conceptual linkage is supposed to be independently intelligible, with the other side derivatively understood in terms of it. So the idea is not to leave out the idea of the right causes, coming to them only as derivative hypotheses of what did the causing on the basis of something we have pinned down as a prior – our blame or our moral sentiments of reaction to the relevant acts.[iv] The blame *is* the blame for something with *that* sort of cause, *but* there is no understanding that sort of cause independent of the blame either. What follows the 'but' in the last sentence rules out the Humean understanding of compatibilism and asserts the quite different Strawsonian one, what *precedes* the 'but', however, equally and precisely rules out the following charge that Baldwin goes on to level against me: ''… Bilgrami would disagree. For he holds that once one accepts that free agency is a normative concept, one must deny that it is a matter of one's action being caused in one way rather than another.'' This is a numbing misrepresentation. Let me just present a *direct quotation* from my book to establish that I am quite innocent of the charge, and move on: ''To reverse the direction would be to imagine that we can have the idea of the relevant reactive attitudes that account for freedom of action without any mention of non-coercive causation. That will seem to many to be just as wrong as it is to imagine –with the [Humean] tradition—that one can have the idea of a non-coercive cause without any mention of the idea of the reactive attitudes. Neither position is

[iv] For my response to McDowell who also misses this point in charging me (see McDowell, ''Reply to Bilgrami' in McDowell and his Critics edited by Cynthia McDonald and Graham McDonald, Blackwell 2006) with excluding metaphysics too much from the notion of freedom, see the discussion of McDowell in Chapter 2.

feasible. So Strawson's main point, cautiously put, is just this last one of the unfeasibility of the tradition.''

6. Baldwin says I make a 'further mistake', this time in my reading of Strawson. I had said, citing various passages from Strawson, that he is complacent in thinking that we cannot raise justificatory questions about whether or not we should exercise reactive attitudes at all. We cannot practically conceive of ourselves as giving them up, Strawson says in those passages, so there cannot be a justificatory question we can raise about them. Baldwin claims that Strawson is in fact providing the justification I think he should be providing for the reactive attitudes when he introduces considerations of 'interpersonal relations' that, as he puts it, '*demand*' the reactive attitudes. Let me quote more fully from Strawson's essay on this very point regarding the relevance of interpersonal relations that Baldwin thinks are so central: ''The human commitment to participation in ordinary inter-personal relationships is, I think, too thoroughgoing and deeply rooted for us to take seriously the thought that a general theoretical conviction might so change our world that, in it, there were no longer any such things as inter-personal relationships as we normally understand them; and being involved in inter-personal relationships as we normally understand them precisely is being exposed to the range of reactive attitudes and feelings that is in question. This, then, is a part of the reply to our question. A sustained objectivity of inter-personal attitude, and the human isolation which that would entail, does not seem to be something of which human beings would be capable,…'' This passage shows very clearly that what Strawson has in mind by the use of the word 'demand' that Baldwin is stressing so much, cannot be interpreted to mean that the reactive attitudes are *justified* by us on grounds of these considerations of interpersonal relations. Rather what 'demand' in that passage makes clear is that the reactive attitudes are *necessarily operative in* inter-personal relations. *There would be no* interpersonal relations without them and, being what we are, we cannot give up these interpersonal relations and the reactive attitudes that are operative in them. So far from showing me to be mistaken, they merely consolidate the point I make about Strawson's complacency in refusing the justificatory question by saying we cannot imagine giving something up.

What do I mean by the 'justificatory' question? I was emphatic over the length of quite a few pages that when I said that we could and should raise the question of justification of the reactive attitudes by further values, I had in mind that there is *no resting point* in the process of justification by values. This was the coherentism of internal justification I presented in the context of this discussion. Even a cursory glance at the passage from Strawson shows that he is going to find any such

justificatory process for the reactive attitudes, anathema. He simply insists that the reactive attitudes *are operative in* these interpersonal relations and there is no question of justifying either the reactive attitudes or the interpersonal relations by appeal to further values. So not only does the passage show that inter-personal relations don't justify the reactive attitudes, it also shows that we must not and need not appeal to any further values to justify either of them, thereby repudiating what I explicitly said I had in mind by raising the question of their justification. A picture (mine) which suggests that if we have interpersonal relations that 'demand' the reactive attitudes in the *non*-justificatory sense of having the reactive attitudes operative in the relations themselves, then we must justify both the interpersonal relations and the reactive attitudes that are operative in them by appeal to further values, is a picture that Strawson finds quite unnecessary. He could not possibly, then, be providing what I think we should be providing in an account of freedom.

Baldwin diagnoses my alleged misreading of Strawson as due to my overemphasis on the first person –what *each* of us would gain in our projects by exercising the reactive attitudes– rather than the emphasis he prefers, on interpersonal relations. But, in fact, one of the values I cite is the value of the inter-personal relation of friendship, which would justify the exercise of the reactive attitudes (if we found further values yet that would justify both friendship and the exercise of the reactive attitudes). It is true that I present the matter as how each of us in the first person would have to value friendship in order for the interpersonal relation of friendship to flourish. But the fact is that for the interpersonal relation of friendship to flourish, each of us *would* have to value friendship and the reactive attitudes among people that underlie friendship. So there ought not to be the deep contrast that he suggests exists between the first person and interpersonal relations. The reason why I stick with a contrast just between the first and third person (a contrast that Baldwin finds too stark because it leaves out these interpersonal relations) is that, for the reason I just gave, there is no understanding the first person without there being interpersonal relations, and in all interpersonal relations, each one of us relates to another with a first personal or *engaged* angle on the other rather than a detached, third personal angle. The idea and the importance of interpersonal relations fall, therefore, *within* a proper understanding of the *first* person perspective, and they are excluded when the focus is on the third person. So understood, there is nothing stark of the sort he finds in my contrast. (I believe this point should also have the effect of showing that the recent stress in moral philosophy on 'the second person' does not introduce any new considerations that are not already there in

the first and third person contrast, but I won't pursue that point here any further.)

7. There is a sympathetic though not detailed gloss from Baldwin on the arguments I give for treating first-order intentional states as commitments, followed by an effort to do my view a favor –showing first, by invoking Moore's paradox, that the paradox establishes that beliefs are commitments, but it cannot establish the same for desires; and then advising me that one may go on to establish that desires are commitments by requiring that they be endorsed (as, say, when we might endorse a disposition we have to do something, as something worth having and doing). This can't really be doing me any favors except in the limiting sense by which favors are done to someone by repeating what he says. On the discussion beginning on p. 212 where I present reasons for treating intentional states as commitments I explicitly say that the term 'desire' is ambiguous between something that is a mere disposition and something that is a commitment and I talk at length about what it would take for it to be a commitment and later in the discussion on p.318, I talk at length of the process in which dispositions get endorsed. (I might add that I don't use Moore's paradox to present any of this because I want to get to the arguments for first-order intentional states being commitments without bringing in *linguistic expressions* of one's second-order beliefs as Moore's paradox does. This is of a piece with why avowals are not central to my argu- ment in the book.)

8. Something that I *had* made central to the book's argument is what I called a 'pincer' strategy (expounded above in my Pre´cis) against the naturalist about intentional states who views them as dispositions rather than as normative states or commitments ('internal oughts'). Baldwin finds the Fregean arm of the pincer inadequate. I had said that if someone thought that beliefs and desires were dispositions, not as a matter of definitional reduction, but as a matter of a posteriori identity, they would be threatened by the Fregean arm of the pincer. Someone can deny the identity without contradicting themselves and one could only account for that fact by positing senses. And I had asked what do these senses express? If they express non-naturalistic properties, the game is over for the naturalist. Baldwin thinks this does not necessarily follow. He says, ''There is a crucial ambiguity in the phrase 'does not express a naturalistic property' between 'does not express a property in naturalistic terms' and 'does not express what is in fact a naturalistic property'. The conclusion follows only if the latter interpretation, that the property is not in fact naturalistic, is assumed.''

I deny this. I think the conclusion follows equally under the former interpretation if there is no eliminating the non-naturalistic terms by

which the property is expressed. Or perhaps I should put the point by saying that I don't think that there is the chasm between the two interpretations that Baldwin thinks there is. Why? Because I assume a Quinean criterion of property existence by which an indispensable element in the concepts or terms which occur in our referential vocabulary, determines our ontology. And until Baldwin shows why that is not the right and honest way to think of property existence, I see no reason why I should not press down with this arm of the pincer. Without some such criterion for property existence, we would likely eventually be led down the path to some version or other of an 'error' theory. That way lies disaster, in my view. I cannot possibly take up the large question as to whether or not an error theory is plausible in a brief response like this, though I do say a little more on the subject when I discuss my notion of 'fool's good' in the 'Reply to Normore' below. I will rest here by saying that Baldwin, by simply assuming without argument that Quine's criterion for property existence is false, has not made his case that this arm of the pincer is ineffective.

Since I will be discussing Baldwin's concluding qualms about the dualism of points of view in sections 6-10 of my ''Reply to Normore'', let me turn to that reply next.

Reply to Calvin Normore

1. Much of Normore's commentary is driven by a dissatisfaction of the radical use to which I put the distinction between the point of view of detachment (what I call the 'third person point of view') and the point of view of practical engagement (what I call the 'first person point of view'). However, he closes his comment with a discussion of my pincer argument whose relation to his objections to the duality of point of view is quite unclear to me. I may have given him the impression that my pincer argument to show that intentional states such as beliefs and desires cannot be reduced to dispositions is in some direct fashion going to establish the point of view dualism that I favor. It could not possibly do that. It only establishes that when normativity is said (as, for instance, by Davidson) to be constitutively relevant to intentional states, those states should not also be thought of as dispositions, as Davidson did. The motivation for my pincer argument was to show that in this regard Davidson had things less right than Kripke who wanted a cleaner break between intentional states and dispositions. None of this had anything directly to do with agency. There is a quite different set of considerations (presented in Chapter 4 and early in Chapter 5, prior to my giving the pincer argument) that try to show how our intentional states, when conceived as commitments and not

dispositions, are essentially integrated with our agency and the fact of our possession of a first person point of view. These considerations invoke the thought experiment (that Normore mentions earlier in his comment) of the exaggerated version of Oblomov, who, I argue, in lacking a first person point of view of practical engagement altogether, also must lack intentionality qua commitment. The pincer argument does not by itself deliver this integration of value, intentionality, and the first person perspective of agency.

Let me, therefore, consider his criticisms of the pincer argument independently and return later to the criticisms he makes earlier of the duality of point of view.

2. He says he is not convinced by the first arm of the pincer (the Moorean arm) because it is 'not of a piece' with Kripke's views on rule-following. Kripke, he says, wouldn't deny that dispositions may be involved in following a rule, he only denies that 'a rule is simply a disposition'. I had been careful to say that I was not concerned to defend Kripke on questions of linguistic meaning and the form of normativity that is supposed to be involved in rules of linguistic usage, but rather I took Kripke's discussion to have a perfectly general relevance (a relevance it obviously has) for the nature of intentionality. When it comes to intentionality, it is my own view that dispositions *are* involved when we act on our beliefs and desires, even when these are conceived as commitments. In fact Normore, earlier in his comment, reports me as saying that a defining condition of commitments is that when we fail to live up to our commitments, we may often want to do better by cultivating the relevant dispositions by which one would be living up to them. This clearly shows that dispositions can be involved. If our dispositions and our commitments are in sync, we would live up to our commitments, if not, we wouldn't. So, my view is quite 'of a piece' with Kripke's. I have just shown how, if we view normativity as relevant to intentionality, then we should take intentional states as being commitments and not simple first-order dispositions, but that does not mean that dispositions are not involved when we live up to our commitments. This is just what Normore reports Kripke as saying about rules: ''A rule is not simply a disposition'', ''but that does not establish that following a rule could not inter alia essentially involve acting on a disposition or being in some other non-normative state.''

He says he is not convinced by the second arm (the Fregean arm) of the pincer either. The second arm of the pincer works only if we accept the Fregean argument and posit a sense for the term 'good' or the terms for intentional states, 'belief', 'desire'. But he sees no reason why a naturalist, who makes an a posteriori identification of good with some natural property x, needs to admit a sense for 'good' to account

for the rationality of someone who denies that ''good is some natural property x'', since (and he presents this as something obvious) we do not need a sense for 'Cicero' to find rational someone who denies that Cicero is Octavian. I am not sure why he thinks that someone who presents the second arm of the pincer as a 'Fregean' arm would assent to this, leave alone take it to be obvious. He seems to think, on the basis of a footnote of mine that he mentions (without quoting it) that I think intentional terms such as 'belief', say, or perhaps value terms such as 'good' need a sense in such contexts, but not proper names such as 'Cicero'. But that footnote merely cites work by Brian Loar and by me on sense and intentionality. It makes no bid to see terms for intentional states as *especially* susceptible to Fregean criticisms of direct referentialist and causal referentialist views, while *immunizing* proper names from such criticisms. My own view is Fregean through and through and proper names are by no means excluded as immune. The familiar strategies that have tried to respond to Fregean criticisms of a Millian view (and its more recent descendants) of proper names are quite unsatisfactory and we do need a sense for 'Cicero' as much as any other term to make sense of the denier of the relevant a posteriori identity. In order to account for the rationality of the denier of the relevant a posteriori identity, these strategies have appealed to such things as diverging causal chains between two different names and the common object to which they refer, or –as in Jerry Fodor[v]– to differential syntactic elements in the language of thought. All these accounts, as I have argued elsewhere[vi], have consequences that fall afoul of a basic desideratum, which is that unless there are good *psychological* grounds (such as, say, our proneness to self-deception) to do so, we should not be said to fail to know our own thoughts. These psychological reasons may be frequently in place and we may therefore often fail to know our thoughts. But at least they are good or appropriate reasons to say we don't know them. You can't deprive people of knowledge of their own thoughts because some Professors of Philosophy (Kripke, Putnam, Fodor) have come up with certain denotational accounts of the terms of our language. In appealing to things (diverging chains of causal relations, syntactical differentials in the language of thought) that are unavailable to the thinker who is denying the a posteriori identity in order to make rational sense of his denial, they fail to meet this basic desideratum. This failure applies across the board whether the denial of identity

---

[v] Fodor, Psychosemantics (MIT Press, 1987).

[vi] See my 'Why Holism is Harmless and Necessary' in Philosophical Perspectives, vol. 12: Language, Mind and Ontology, Blackwell 1998. For a more extensive treatment of these issues, see my book Belief and Meaning (Blackwell, 1992)

involves proper names such as ‘Cicero’, or terms for intentional states, such as ‘belief that p’, or value terms such as ‘good’.

3. Normore cites some *additional* considerations I give for not leaving sense out for the last of these terms, ‘good’. He agrees with me that if after reference-fixing was done, sense dropped out altogether from our understanding of a term such as ‘good’, then it would be possible that all the judgments we have ever made about good could be quite wrong; it is possible, in other words, that we have all along been chasing what I called ‘fool’s good’. I had claimed that this is a reductio ad absurdum of the view that leaves sense out of our understanding of ‘good’. If someone wanted to insist that it *really* was the case that we have had ‘good’ completely wrong in every use of it in all the value judgments that we have ever made, then we should be quite brazen and say, ‘‘Well, fool’s good will serve just fine for the purposes of living our normative lives of mind and action.’’ He thinks that it is easier to imagine such a reductio for terms such as ‘heat’ but not for ‘good’ since we disagree so much about good. He describes it as complacence on my part to assert a certainty that we have not made wholesale errors of identifying the extension of a term such as ‘good’ when there is so much disagreement about what is and is not good.

I don’t deny that there is a great deal of disagreement over what is and is not good. I don’t deny either that many particular applications we have made of the word ‘good’ may be wrong. They could hardly fail to be, since when there is the disagreement about good that Normore stresses, there may well be two contradictory assertions being made about what is good –and if so, only one of the assertions can be right. But I think it is a *non sequitur* to go from conceding all this to saying that it might be all right to say that all uses of the word ‘good’ that have ever been made in value judgments have been wrong and we should admit that we might in our mind and action be living in a normative void. And it can’t be a complacence to resist falling into a fallacy. Why do I say that it is a non sequitur? Because unlike, ‘fool’s gold’, which at least has the semblance of something we can plausibly posit, ‘fool’s good’ (in my somewhat ostentatious appeal to such a thing) is not really a seriously entertainable idea. Why not? Because our use of the term ‘good’ is governed by certain features that are not dismissible as superficial properties, such as ‘the yellow metal that glistens under light’, etc. that *are* so dismissible in the case of ‘gold’. There is no space to discuss those features of ‘good’ here, but the point can nevertheless be briskly conveyed. Given the kind of normative concept it conveys, ‘good’ is governed by such things as universalizability, rankability, aggregateability, possession of motivational force, etc. (Someone may quarrel with the inclusion of one or other of these

features, but not all.) And if our judgments with the use of the term 'good' are such that they conform to these features, then either we should say that the idea that there is 'fool's good' is not a seriously entertainable idea as fool's gold *migh*t be, or –if someone high handedly insists that we rigorously see through the consequences of a certain theory of reference for all terms and insists, in turn, that there might be fool's good too, then— we should say that fool's good because it also has conformed to these features in our judgments is a perfectly serviceable substitute for the normative concept that the term 'good' was intended to convey. If this means one is cleaving to the notion that good must have a sense, so be it. That was, anyway, what the reductio was intuitively supposed to achieve.

In order to say any or all of this, I need not assume that there is the kind of convergence and agreement in our use of normative terms such as 'good' as there is in terms such as 'heat'. However extensive that non-convergence and disagreement is, the underlying point is that a certain central normative concept –with certain governing features– that we take ourselves to be expressing with the term 'good' is not an eliminable normative concept; and if some theories of reference want to force on us a certain outcome, viz., that we may have never made correct judgments with the term 'good' and have been expressing a chi- merical concept in our use of that term (not good, but fool's good), then there is nothing to say but that that allegedly chimerical concept is *the* ineliminable normative concept that governs our minds and actions.

4. Before I come finally to a defense against criticisms from both commentators on the duality of points of view, let me just register agreement with an excellently perceptive point that Normore makes, which I wish I had made more conspicuously than in a footnote, since not to have it conspicuous may have the potential to mislead a reader about my views on the nature of commitments. He points out quite correctly that beliefs that a subject is committed *to* may not be transparent to the subject. And, as I say (see footnote 7 on p. 371–2) I do not call what a subject is committed *to*, her 'commitments'. A belief of mine (thought of as a commitment) generates various other things that I am committed to, but these latter are not my 'commitments', until I acknowledge them. (Thus, in his discussion of this point, what Normore calls 'reasons' may just be –I am not sure– synonymous with what I call commitments.) This, as I say in that footnote, distinguishes my use of the word 'commitment' from other standard uses in the philosophical literature and makes it much more restrictive. Indeed distinguishing my idea of commitment from others in the recent literature is the point of that footnote. For me, the notion of commitment in the study of intentionality is motivated by a framework in which not

merely normativity, but *agency* is central to intentionality. That forces a restriction on what gets to count as a commitment. I won't get into a detailed discussion of this here, except to say just this: there could be no justification for indiscriminately making *all* of what my beliefs commit me to, the target of the reactive attitudes that are directed at me.

5. Normore begins his criticisms of the claim of a duality of point of view by speaking in support of John McDowell's view that there is a univocal notion of 'cause' that holds both in the explanation of action and in natural scientific explanation. He says, following McDowell, that it is a distortion, owing to Hume, to think that this might not be so because one thinks, wrongly, that causality implies laws. Normore goes well beyond McDowell in claiming that laws should not be given the central place some philosophers think they have in the natural sciences. I will not address this last point since there is something underdescribed in Normore's statement of his opinions on this subject. (For instance, the example in –and the authority of– writings by Anscombe and Feynman that he cites are too cryptic for me to be able to address since I am not sure exactly what I would be addressing.) McDowell himself understands causality as deployed in natural science to be embedded in the aspiration to laws. He may acknowledge that in some recent understanding of explanations in biology as, for instance, when the notion of a biological function is on centre-stage, the nexus of cause and law may fall away, but even that suggests only that the aspiration to law is not comprehensive, not that it is not exemplary. What McDowell says is that even if that is so, and even if explanation that makes sense of human action has no such exemplarity in mind because it uniquely appeals to normative considerations of rationality, that does not mean that there is not a univocal notion of cause in both natural scientific and rationalizing explanation, as I had said there was not.

Why had I said this?

It was important to my argument that intentional states, even if thought of as commitments and not dispositions, should be such that they make a difference to the world and are not epiphenomenal. Causality *is* in play in the relation between intentional states (qua commitments) and actions. It makes perfectly good sense to say: ''Her commitment to… caused her to…''

But it was equally important in how I had drawn the contrast between commitments (thought of as normative states in the manner that I had characterized them) and dispositions (thought of naturalistically, in the sense that they are studied by the natural sciences) that it resulted in the following crucial difference in the causal relations in which each of these stand to human behavior. When one sometimes finds that the commitment marked by the first ellipsis above does *not*

result in the relevant action that is marked by the second ellipsis, we do not think (as we would with the causal claims in natural science such as those involving dispositions) that there is an impending *refutation* of the connection one had made between that cause and that effect. There is no such threat of refutation and all that the non-occurrence of the relevant action generates is an attempt on the part of the possessor of the commitment to try and do better by way of living up to it. This is, I claimed, a very distinctive understanding of the notion of cause and there was no point in yoking it together by lexical stipulation as being the same notion that was operative in the natural sciences, where such an impending refutation of a statement describing causal relations *does* loom if there are failures in the occurrence of the events marked by the second term in those statements. Thus, when Normore writes: ''Once one… [allows that] like causes need not have like effects, that does not mean the notion of cause has changed'', I want to ask: why not? To my ear it comes off as mere assertion on his part, *the more so* now that I have presented this quite distinctive property that holds of those causes that are intentional states, thought of as commitments.

6. So far, I've discussed the univocality of the notion of cause as a self-standing question. But, Normore gets down directly to its deeper implications when he says: ''If there is a notion of cause which is neutral between intentional and non-intentional cases, then there is a way in which as agents we can affect the non-intentional world. If there is not, if all agential activity is inside the realm of the normative, then we need a story about the connection between an action (closing the door) and the changes in relative position of various pieces of material before we have an understanding of how our action relates to the non-intentional world.''

This raises the heart of the difficulty that Baldwin also wants to raise about the view that I describe with my recklessly adopted expression 'dualism' of point of view.

Ever since the powerful critiques of Cartesian philosophy in the last century, 'dualism' has been a word of opprobrium. Philosophers today make distinctions, they make differentiations, but they resist elevating these into a dualism, if they can help it, and they certainly avoid *calling* them 'dualisms'. The passage I have just quoted has all the marks of a recoil from Cartesian dualism. We are urged to relate elements in the normative, agentive point of view with 'changes in the relative position of various pieces of material'.

No qualm is induced in me by this demand. It is not an issue for my view, and it should not be directed against my view. The agentive or first person point of view unproblematically includes within its elements

'various pieces of material' and the changes that they may undergo. There is nothing, therefore, to do by way of bringing these things together. The normative, the agentive, the intentional, are not relegated to a gratuitous metaphysics of a noumenal realm of pure practical reason. Nor do the normative and its contrast coincide with the Cartesian contrast between the mental and the material. We can scramble these two contrasts at least to the extent, as I said, of having the normative contain the material. Mine is a dualism of point of view, not substance.

The fault-line in Descartes' philosophy was a conflation –to take the subjectivity of mind (which since his time has tended to focus on the qualitative side of subjectivity, not the agentive side which is our concern) that he establishes in the first two 'Meditations' and *spoil* it in the subsequent 'Meditations' by viewing mind as a substance, thus making it into something that falls within the *objective* point of view. The objective point of view can, thus, have within its purview not only matter, but also this allegedly immaterial *substance*, 'mind', which now, being substance, is apprehended within a third person point of view of detached objectivity. But equally this allows us, shedding or reversing such a Cartesian picture, to say this: the first person point of view of subjectivity can include within it something that is material, so long as it is viewed from the point of view of agency and not detachment. I can see the (liquid) matter of the Colorado River as H_2O but I can also see it as having an ecological value. It's not as if I fail to see it as (liquid) matter in doing so. It's rather that as matter, it is shot through with the value. So also, when I, *as an agent*, act upon a door and shut it, it is not as if I have abandoned the pieces of material and gone off to some other ethereal realm. It is the material door that I have acted upon. Normore's anxiety about bringing together the normative with the material puts no pressure on my views. The material is, without any strain of metaphysical effort, included within the normative point of view. Nothing Cartesian contaminates my dualism.

7. There is hereabouts a point of further contrasting interest with Descartes. Cartesian dualism, as I have been saying, is a distinction made *within* the point of view of objectivity. Once it is conceived as a substance, and not a point of view, mind, like matter, is relegated to one of two dichotomous categories, *both* to be apprehended within the third person point of view. The dualism of agency and detachment, on the other hand, is such that *one cannot be in a richer point of view that straddles both*. One is always in one or other point of view. Normore writes to say that this is wrong. Straddling should be possible and is frequent. Both points of view are abstractions from something richer, he says, and calls this richer point of view 'that of a participant in the world'. Here is how he argues for this conclusion. I had asked the

reader to imagine a superlatively exaggerated version of Goncharov's character Oblomov who has thoughts entirely in the mode of the third person. Normore asks us to imagine the mirror image of such a character, one who has only thoughts in the first person. He is agnostic as to whether or not we can imagine such a subject, but he is certain that if we could, we would have imagined a subject whose first person point of view is not at all like ours. This 'encourages him', he says, to conclude that both our first and our third person points of view are abstractions from something richer. I don't feel the tug of any such encouragement. I think the facts, as he describes them, underdetermine encouragements on the matter. Why should it not be just as easy to conclude from the fact that our first person point of view is quite unlike that of the imagined counter-Oblomov because it sits side by side with a third person point of view that is missing in the counter-Oblomov, that we have no other richer point of view that straddles them both, that we have only these two points of view, which crowd each other out, such that one is either in one or the other, and never straddling both?

Though I doubt it, it is just possible that Normore has been misled by the details of some of this talk of how a subject is always in one or other of the detached and engaged points of view, to think that I have argued for something stronger, at any rate, something other than I have. If that is so, then there may be more agreement between us on this topic than the tone and tendency of his comments indicate. In my dialectic I had started with the familiar distinction between a subject intending to do something and predicting that she would do something, saying that the one crowded the other out. I had seen this as reflecting something of greater generality: intending and predicting were special instances of something more general, the points of view of agency and detachment, two different perspectives we could take on ourselves, each of which crowded the other out. And finally, I had argued that we could extend this to two different points of view of engagement and detachment we could take, not just on ourselves, but on the world. Now, consider the detached point of view one might take on oneself, when someone says, ''I predict that I will do…'' I had pointed out that the first ''I'' marks a subject in the first person perspective, the subject as agent. It is only the second ''I'' that marks the subject as object. So, in one perfectly innocuous sense I had myself claimed that the two points of view on myself are points of view within engagement or agency or 'participation in the world' to use Normore's expression. The first occurrence of 'I' both in ''I intend that I…'' and ''I predict that I…'' is the 'I' of *agency*, after all. This holds also of the two points of view on the world. Even when I am studying the world in a detached

way (say, as a natural scientist), I am still *an agent* who is studying it. It is my *angle on the world* that is detached. If this is all that he means by how the two points of view are points of view within a participant's stance, I myself had also meant (and said) just that. I suspect, however, that it isn't all he means – first, because his rhetoric of the first and third person stances as being '*abstractions from*' a richer stance is quite the wrong way to describe the fact that it is as agents that we take both the first and third person stances; and second, because there is nothing in any of this to undermine my point that we are always either taking a detached or an engaged angle on ourselves and the world because the one crowds the other out. Yes, of course, it is *as agents*, that we are in one or other stance on the world or on ourselves, each of which crowds the other out. But the fact that each *does* crowd the other out does mean that we cannot straddle both points of view. It is this impossibility of straddling both in a richer point of view which is neither one of these points of view on the world or on ourselves, that I am insisting on and on which Normore explicitly opposes me. But, as I said, I see no health in the argument that he gives for his opposition, when he appeals to the difference between the counter-Oblomov's first person point of view and ours.

8. At one stage in the discussion, he invokes Gibson's notion of 'affordances' in the perceptible world, which are defined in terms of an animal's possibilities of actions. He asks whether an animal psychologist who studies affordances in a detached way is seeing the same property in the world or a different one from the one we see as agents, when we see it as calling upon us to action. The answer is we see it as a quite *different* property (though *both* angles may involve 'pieces of material', and so no Cartesian dualism is generated by this affirmative answer to the question.) So also a chemist may see H_2O in the same place where someone else sees an affordance, an *opportunity* to quench his thirst. They see different properties in the same place. That simply follows from the dualism of point of view. From one point of view one does *not* see an *opportunity* as one does from the other. Normore simply plunks it down that it is not a different property in the world without giving any reason for doing so. Indeed, one may say that the chemist may herself see the affordance or a calling to action (the opportunity to quench her thirst) over and above seeing a chemical compound, *but not without defecting to another act of perception*, a first personal perception of the world as offering her an opportunity.

It is at this or on this point that both Normore and Baldwin converge on the same criticism.

In order to convey how an agent can go from seeing something in one (detached, third person) to the other (practically engaged, first

person) point of view, I had given examples in my book of how someone may see something as a chemical compound or an opportunity to quench one's thirst, a meteorological perturbation or a threat to one's thatched dwelling, an average daily ingestion of a certain caloric count or a case of need, malnutrition starvation… Both Baldwin and Normore rightly point out, as I myself had, that our *practical* deliberations of *engagement* (say, before giving money to Oxfam), may involve us in some *detached* theoretical calculations of caloric counts (say, of a peasant population in Africa).

But then: ''That is absurd'', Baldwin says, of the idea that each time we act on our practical deliberations, we have to produce a translation or re-interpretation of the vocabulary of 'caloric counts' into the vocabulary of value, such as 'need' or 'malnutrition' or 'starvation'. Well, that would indeed be absurd. But the only person who has put such an absurd idea into the air (or typed it on a keyboard) is Baldwin. What I said was that we defect from the deliverance of a third personal detached calculation to another kind of perception and see that fragment of the world as making a normative demand on us and engaging our *first* person point of view. The claim is a *philosophical* claim about perspectives or points of view, it is *not* a linguistic or grammatical claim. I said so in exactly those words in the book, pointing out that the same linguistic term 'I' can be used to talk of oneself as an agent and of oneself as an object of one's prediction, that is, the same word may surface to describe quite different elements in the two different points of view. So also, one need make no translations whatever between the vocabulary of caloric counts and the vocabulary of needs. One may in fact not even *have the vocabulary* to make the translations (leave alone make them) in order to defect from one point of view to another. All that is being claimed is that a subject, in fulfilling what Normore, following Gibson, calls 'the possibilities of action', sees things from the point of view of engagement, even if, in his deliberations, he had started out with some detached calculations. I, as philosopher, may try to convey this duality of point of view by introducing the non-evaluative and evaluative vocabulary I possess ('caloric counts', 'malnutrition', respectively) to make things vivid. But the philosophical point itself is not to be laid over with some linguistic demand on the subject in question that she provide a translation or re-interpretation before she fulfils, what Gibson calls the 'possibilities of action'. Words will fall where they may on both sides of the first/third person divide and for some subjects, the same word may fall on either side of the divide. I have given the example of 'I'. The same is true of 'need and 'caloric count'. I can think or say: ''This number I have assigned to this peasant's caloric count amounts to a need, a case of malnutrition.''

And such a thought or statement, put in these words, may well be on the side of detachment, despite containing the evaluative *words* 'need' and 'malnutrition' over and above the vocabulary of 'caloric counts'. The words are not the important thing; it is the detachment, the third person point of view. All I required was that one has to defect from this to a point of view of engagement to see that fragment of the world as making a normative demand that leads to the fulfillment of 'the possibilities of actions'. Equally, someone may describe what makes the calling on her 'possibilities for action' (i.e., the normative demand from some fragment of the world on her first person point of view that leads to her action) in terms of the vocabulary of caloric counts: ''Let me help these peasants, they have an average caloric count of….'' There is no need to reinterpret this into value *talk.* All that is needed is the exercise of the normative, agentive, first personal stance, now no longer the calculating, detached stance.

The point should be familiar and uncontroversial from other points that we often make, such as this: If I think of a desire in the third person mode ''This is desired by me'', it has no motivational force of agency for me. By contrast, if I experience it in the first person mode, ''This is desirable'', it does have such a force. I do no more than deploy points of this kind to make the claim of a duality of perspectives. Baldwin himself begins his essay by characterizing my point- of-view dualism in familiar and uncontroversial (and accurate) terms but then succeeds by the end in putting a gloss on it that makes it sound, to use his own word, 'absurd'. The construction of an absurdity is his doing, not mine.

9. I am charged by him with Sartrean excess on the subject, even though it is Spinoza I allude to in distinguishing between intention and prediction, and the more general first and third person points of view. Baldwin thinks that Spinoza, being a monist, is the wrong hero for me. (I had myself pointed out that I do not follow Spinoza to the point where he thinks it necessary to dissolve the duality of point of view and proclaim a monism). In any case, I think the real excess that Sartre himself should be charged with is not the making of a point that should be obvious (a distinction or 'duality' between first and third person points of view), but the totalization of one of these points of view, the first person point of view, a totalization I strenuously disavow. I have no phobia against detached third personal angles on the world (as is found, say, in science) or on ourselves (as is found, say, in psychoanalysis), the last of which is said by Sartre to generate 'bad faith'. For this reason I am quite the wrong target for Baldwin's attack, when he says that a focused dose of 'inter-personal' relations would undermine my excessive stress on the first person. As I said earlier, interpersonal

relations are very much part of what is meant by me as 'the engagement' that goes into the first person point of view. It does not have to be *imported* into my scheme of things in order to improve it, it is an essential part of my scheme of things, which is unimprovable in this respect. Inter-personal relations do nothing, therefore, to undermine the duality of point of view that characterizes my scheme of things.

10. I suspect that what really causes such indignation (''This is astonishing'', ''absurd'', ''a dead end'') in Baldwin (Normore's doubts on the matter are expressed more soberly) is that my point-of-view 'dualism' of agency and detachment is said to render unassessable the thesis of even the weak dependency relation of value properties on natural properties (as the natural sciences study them) that we have come to call 'supervenience'. I had argued in the book that such a relation seemed plausible to many while the notion of value was not seen as deeply integrated with the notion of agency in the way that I try to present in Chapters 4 and 5 of the book. The 'defections', which I have recurrently spoken of, that we must frequently make between the points of view of detachment and engagement, were a provocative way of putting into doubt the thesis of supervenience – not its *truth*, but its assessability as a thesis. I certainly would not want to *deny* the thesis. Denials of it come off as 'absurd' because they assume that it is something assessable, i.e., deniable or assertible, in the first place, and once you assume that, it seems extreme to take the side of denying it. But if value is deeply integrated with agency and engagement, as I try to show in the book, and something like 'defections' captures how we move from detachment to engagement and vice versa, the thesis of the supervenience of value on natural facts (as the natural sciences study them, i.e., with detachment) is no more assessable than the idea that duck-facts are supervenient on rabbit-facts. Indeed it is the idea that this might be assessable which is 'absurd'.

In the face of failures of attempts to *reduce* agency and value to natural facts (as studied by the natural sciences), the much weaker dependency relation of supervenience may seem to be our last resort in preventing the disunification of nature as containing elements that stand so deeply at odds with one another as value and agency do with the natural facts that are studied by natural science. Does the duality of point of view between agency and detachment have implications that render nature unredeemably disunified? I don't see that it does. It does not render false or unassessable claims such as these: ''We are enabled to have agency (in the normative and self-knowledgeable form that my book has elaborated) by having evolved as we have into the kind of creatures we are''; or ''We could not have agency in this normative and self-knowledgeable sense, if we did not have a large enough brain of

the kind we do.'' These are not very exact things to say, but they say enough to show that such an integration of value and agency does not present itself as a disruption of the natural world. There may indeed be further (even perhaps stronger) things to say that would show nature to be unified despite the implications of the duality of points of view. But that only means that there is more philosophy to be done, not that we should assume forms of unification that can't so much as be assessed.