TRANSPARENCY, RESPONSIBILITY, AND SELF-KNOWLEDGE

Henry Jackman
York University

There is currently a large literature on the topic of self-knowledge that focuses on arguments of the following general type:

1. The content of what we think is determined by factors external to us.
2. There are aspects of these external factors that are not transparent to us.

---

3. Therefore, there are aspects of the content of what we think that are not transparent to us.[1]

For instance, if I don't know anything about chemistry, and what I mean by "water" is a function of the chemical structure of the stuff I call "water" in my environment, then there will be facts about what I mean by "water" that are not known to me.

Given the prevalence of such arguments in the literature, it is noteworthy that Akeel Bilgrami's recent (2006) book on self-knowledge, *Self-Knowledge and Resentment*, manages to proceed for almost 400 pages as if such arguments drawing out the tensions between externalism and self-knowledge did not exist. It certainly isn't as if Bilgrami is unaware of this literature, and early versions of the ideas contained in his book can be found in articles that are explicitly situated in the context of this currently more mainstream debate about externalism and self-knowledge.[2]

Bilgrami's silence on this topic can be explained, I believe, in terms of his conviction that arguments like that embodied in 1-3 rest of a fundamental misunderstanding of the nature of externalism, a misunderstanding he takes himself to have already cleared up in his earlier book (1992a), *Belief and Meaning*.[3] Bilgrami there argues that, according to "orthodox" versions of externalism, the contents of our thoughts are characterized in terms of things we may not be aware of at all (such as the physical structure of our environment for the 'causal' externalist, or the usage of our community for the 'social' externalist), and that it is precisely this independence that leads to arguments like 1-3 above.

By contrast, Bilgrami (1992a, p. 5) argues that accommodating self-knowledge is a fundamental desideratum of accounts of mental content, and

that any externalist theory of content should be structured by the following constraint:

> (C) When fixing an externally determined concept of an agent, one must do so by looking to indexically formulated utterances of the agent which express indexical contents containing that concept and then *picking that external determinant for the concept which is in consonance with the other contents which have been fixed for the agent*.

So, for instance, while we may attribute thoughts about water to a chemically ignorant speaker because there is, in fact, water in his environment, the contents of his thoughts must be characterized in terms of, say, "water," "the clear colorless liquid that I drink every day" or some other characterization that he would recognize, rather than, say, "H_2O," or some other characterization that he would not.

Bilgrami motivates (C) with the thought that we are rationally responsible for our beliefs, and without (C) beliefs would not be the sorts of things whose contents would bear rational relations to each other. Without (C), we could make ascriptions such as "Bert believes that he has a disease of the joints in his thigh," "Luthor believes that Superman is not Superman," or any other of a number of patently inconsistent beliefs that would seem incompatible with the rationality of the believer. If we endorse (C), by contrast, we can say, for instance, that while Locke may have believed that there was water in his bathtub, he did not believe that there was H_2O in his bathtub, since characterizing water as "H_2O" in our attributing a belief to Locke would violate (C). According to Bilgrami, orthodox externalists are not in a position to say this, which is why they are vulnerable to arguments like 1-3. By endorsing (C), Bilgrami insists that the external factors in premise (1) be characterized in ways that are transparent to the agent, so that premise (2) no longer holds for them.[4]

Bilgrami thus already has an externalist conception of content that is friendly to self-knowledge in place with *Belief and Meaning*. Content is still determined by factors that are external to us, but those external factors must be mediated by our conceptions of them in a way that blocks arguments like 1-3.

However, while the version of externalism defended in *Belief and Meaning* is friendly to self-knowledge, it doesn't guarantee that we have such self-knowledge. Constraint (C) simply blocks a familiar argument that purport to show that externalism and self-knowledge are *incompatible*, but merely showing that self-knowledge is *compatible* with externalism doesn't show that we *must* have it. On its own, (C) won't require that my merely believing that, say, water is dangerous entails that I know that I do, since the

fact that the belief is characterized in accordance with (C) doesn't in any way entail that I will know that I have it. A belief that I was unaware of could still have its contents characterized in ways consonant with other beliefs that I had. One wants more than just a defense of the claim that self-knowledge is *possible*, one wants an account that explains why we *actually* have it, and providing such a positive account is one of the primary goals of *Self-Knowledge and Resentment*.[5]

Bilgrami thinks that the relevant notion of self-knowledge can be captured in terms of our knowledge of our own thoughts being both "authoritative" and "transparent," with this Transparency and Authority being captured by the following two conditionals:

> **Transparency (T):** It is a presumption that: If S desires (believes…) that *p*, then S believes that she desires (believes...) that *p*.
>
> **Authority (A):** It is a presumption that: If S believes that she desires (believes...) that *p*, then she desires (believes...) that *p*. (Bilgrami, 2006, p. 89)

Authority is far from uncontroversial. However, I won't be discussing it further here, since the doubts about Authority don't arise naturally from tensions with semantic externalism in the way that doubts about Transparency do. A chemically ignorant speaker may raise worries that he believes that H_2O is wet without believing that he does, but there are no worries that he would believe that he believes that water is wet without actually believing it.[6]

Transparency is our main concern here, and Bilgrami believes that something like Transparency must be true because, without it, we couldn't be held *responsible* for the actions that our beliefs and desires cause. As he puts it:

> It is a conspicuous fact about the notion of agency… that it takes for granted self-knowledge. Actions do not *justifiably*… get counted as responsible if the actor does not know that she has acted in that way…. The claim, then, is that these attitudes and practices are not justifiable when they target actions which are not self-known by the agents who perform them. And from this emerges another related point. These actions are not justifiable targets of such attitudes or practices, not only when the actions themselves are not self-known to the agent, but also if the intentional states which rationalize the actions are not self-known.… If this is, in first approximation, right, then we can say that to the extent that internal states fall within the region of

> responsibility, that is, to the extent that they are tied to responsible action, then there must be self-knowledge of them. In other words, transparency of intentional states is established so long as the intentional states have such ties to responsible action. (Bilgrami, 2006, p. 93)[7]

Self-Knowledge is thus tied to our conception of ourselves as responsible for our actions, and such a self-conception is, according to Bilgrami, not negotiable. We must see ourselves as responsible for our actions, and thus must see the contents of our thoughts as transparent.

In short, Bilgrami consciously follows the pattern of Strawson's (1974) "Freedom and Resentment" in presenting an argument something like the following:

1. Our reactive attitudes require that we know what we are doing in order to be justified.
2. Knowing what we are doing requires knowing what we are thinking.
3. Our reactive attitudes require knowing what we are thinking. (from 1 & 2)
4. Our reactive attitudes are a non-negotiable part of our lives.
5. Our knowledge of our own thoughts is non-negotiable. (from 3 & 4)

Arguments of this "Strawsonian" form are obviously controversial, and some have characterized them as little more than a sophisticated presentation of wishful thinking. Bilgrami spends most of the second chapter of his book defending the legitimacy of this sort of reasoning, and in what follows I will simply grant that his defense of the argument pattern of "Freedom and Resentment" is successful. I will focus instead on the question of *even if* such arguments are successful, how robust of conception of self-knowledge they support.

Concerns that the conception of self-knowledge that follows from Bilgrami's arguments might not be quite as robust as he supposes follow from the following consideration: if the justification of (T) is tied to our notion of responsibility in a way that limits Bilgrami's defense of self-knowledge to the domain of ascriptions of responsibility, then such restrictions should be built into (T) itself. Indeed, this is precisely what Bilgrami does in the following reformulations of the transparency constraint.

> (T): Given agency, if S desires (believes) that *p*, then S believes that she desires (believes) that *p*. (Bilgrami, 2006, pp. 119, 138)

Of course, "Given agency" does a lot of work here, and Bilgrami (2006, p. 119) admits that "Fully (re)written out, the conditional would read":

> (T)* To the extent that an intentional state is part of a rationalization (or potential rationalization) of an action 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.

However, when transparency is understood this way, then it seems that the required scope of self-knowledge is not as extensive as it might have seemed in *Belief and Meaning*. In particular, failures of self-knowledge are possible on such an account provided that the failures in questions are not relevant to our justified reactive attitudes.[8]

Bilgrami stresses this himself when he argues, for instance, that the fact that certain 'subconscious' or 'repressed' beliefs of the sort associated with Freudian psychoanalysis are not transparent to us should not be viewed as counting against the truth of (T) because, while such beliefs may exist and cause us to act in certain ways, we are not morally blameworthy for the actions that are so caused.[9] The beliefs involved, even if they manage to cause some of our actions, do not enter into our conscious practical reasoning.

However, if we only need have knowledge of our intentional states "to the extent" that they are (or produce actions) that are objects of justifiable reactive attitudes,[10] then there is a good sense that our mental states need not be *completely* transparent to us. Much of their content can be opaque, provided, of course, that such aspects of their content aren't the objects of our reactive attitudes.

For instance, it has been suggested that externalism entails that we lack knowledge of 'comparative' content. That is, while we know that we believe that water is wet, we may not know whether that involves believing that H_2O is wet or XYZ (a complex substance superficially indistinguishable from H_2O) is wet.[11] Bilgrami gives us no reason to think that the difference between *these* two contents need be transparent to us. After all, the content of our thoughts need only be transparent to us "to the extent" that they contribute to the moral evaluation of the resulting actions, and for the most part, the difference between these two contents is morally irrelevant. There is no moral difference between planning on drowning a puppy in a tub of H_2O and planning on drowning a puppy in a tub of XYZ, and so the agent's ignorance of the difference between these two thoughts does not make his action any less culpable. The morally relevant features of the plan, that it involves submersing a puppy in a liquid that will drown it, remain the same. If some aspect of a particular thought's content is morally irrelevant, then Bilgrami's account gives us no reason to think that it need be transparent to us.

Bilgrami tends to focus on how the Strawsonian picture allows that *particular thoughts* (e.g. subconscious ones) need not be transparent to us, but the view also seems to entail that *particular aspects of our thoughts*, even of our conscious thoughts, need not be transparent to us either. Indeed, the proposed connection between belief, self-knowledge, and action suggests that the requirements of self-knowledge, and its connection to rationality, are more limited than Bilgrami originally suggested. In particular, it may turn out that constraint (C) is not a constraint on belief content *per se*, but rather just on a particular aspect of it.

We do have a rational responsibility for our beliefs, but it need not follow that their content must be exclusively understood in terms that are friendly to such responsibility. To see why, consider the analogy with action, and our responsibility for it.[12]

We can, for instance, distinguish:

1a. The things we do.

from

2a. Descriptions under which we are morally blameworthy for what we do.

To take a familiar example, if, unknown to me, someone rewires my light-switch so that turning it on triggers a bomb that kills a room full of innocent people, then if I turn on the light when I come home, I will have also triggered an explosion killing a room full of people. Still, while triggering the explosion is something I did, I'm certainly not *blameworthy* for it (though I may in some sense be *responsible* for it).[13] Our actions have properties that we are not aware of, and are in this sense, not *transparent* to us. However, we can only be blamed morally for those aspects that are transparent to us.[14]

In much the same way, we can distinguish:

1b. The things we say or believe.

from

2b. Descriptions under which we are rationally blameworthy for what we say or believe.

So, if I am unaware of any of the relevant facts about chemistry, I may believe that water is not H_2O, and so could be correctly described as believing the contradictory content that H_2O is not H_2O. However, I am not *irrational*, and I am not rationally *blameworthy* for that belief, since it is presented under a description that plays no more of a role in my theoretical reasoning than the description "triggering the explosion" played a role in my practical reasoning when I turned on the light switch. (I may, however, still be rationally *responsible* for this belief, in that I am responsible for its consequences and rationally should give it up when confronted with these.)

However, the fact that the chemically ignorant person isn't rationally culpable for believing that H_2O isn't H_2O, no more entails that it isn't something that they *believe* than does the fact that I wouldn't be morally culpable for triggering the explosion entails that it wasn't something that I *did*. Responsibility doesn't entail complete transparency with our actions, and there is no reason to think that it need do so with our beliefs.

It might seem, then, that Bilgrami's constraint (C), rather than being a constraint on thought content *per se*, is only a constraint on though ascriptions involved in evaluating others as rational agents. This is, obviously, an important (if not essential) part of our interpretive practice, but this is not the *only* role belief ascriptions can play. We may simply be interested in the truth of the interpretee's beliefs, or in which items in their environment they are related to, and these interests may be best satisfied by ascriptions that specify the interpetee's beliefs in terms that they would not recognize. The claim that Oedipus intends to marry his mother may not capture the content of his belief in a way that preserves his rationality, but it captures an important fact about Oedipus that may be of independent importance to the claim's audience.

If this is the case, then Bilgrami's original constraint (C) would be better formulated as:

> (C*) *To the extent that an intentional state is part of a rationalization (or potential rationalization) of an agent's actions and inferences*, to that extent one must fix the externally determined concepts of an agent by looking to indexically formulated utterances of the agent which express indexical contents containing that concept and then *picking that external determinant for the concept which is in consonance with the other contents which have been fixed for the agent*.[15]

However, while (C*) may be true, replacing (C) with it closes much of the distance between Bilgrami's externalism and the more "orthodox" externalisms that he associates with writers like Putnam and Burge. While Bilgrami originally argued that the sorts of ascriptions that followed from such views (and which didn't respect (C)) were *false*, it now turns out that they simply may be, while true, inappropriate for rationalizing explanations, and it isn't clear why the orthodox externalist need have any problems taking that sort of qualification on board.[16]

Of course, this move from (C) to (C*) would also entail that our initial argument about the tension between externalism and self-knowledge may still have some go. However, if the aspects of thought content that we may not know are relevant neither to our rationality nor our morality, then a lack of transparency in this department is something that we can live with. Just as

there are facts about what we do that we are unaware of, there are facts about/aspects of what we believe or say that we are unaware of as well.

It might seem, then, that understanding why self-knowledge is important also make it clear which types of self-knowledge are *not* important, and thus why we can live with the gaps in self-knowledge suggested by arguments like 1-3.[17]

Notes

[1] See, for instance, the essays in Ludlow & Martin (1998), Wright, Smith, & Macdonald (1998) and Nuccetelli (2003), as well as the literature cited therein.

[2] As titles of earlier papers like "Can Externalism be Reconciled with Self Knowledge" (Bilgrami, 1992b) should make clear.

[3] These issues are also addressed explicitly in Bilgrami (1992b).

[4] Or, conversely, it requires that the non-transparent external factors mentioned in premise (2) not enter into the content ascriptions needed to make premise (1) true.

[5] Though the germ of this argument goes back as far as the appendix to Bilgrami (1992a).

[6] Of course, such a speaker may not be authoritative about what he *doesn't* believe (e.g. he may claim that he doesn't believe that H_2O is wet even though he actually does), but this lack of 'negative' authority is still compatible with the truth of (A), which is only about the claims about which mental states we *do* have.

[7] For an earlier version of this claim, see Bilgrami (1999, p. 217):

> It is a conspicuous fact about responsibility…. that it takes for granted self-knowledge. Actions don't … get counted as responsible if the actor does not know that she has acted in a certain way. If this is right, if self-knowledge is a necessary condition for responsibility, then we can say that to the extent that intentional states are in the realm of responsibility, so long as they are tied to responsible action, then there must be self-knowledge of them. That is, transparency of intentional states is established so long as the intentional states are in the region of responsibility.

[8] Which may include judgments of irrationality as well as moral blame.

[9] For a fuller discussion of psychoanalysis, see the appendix to Bilgrami (2006).

[10] See also, "The connection remains, only now it remains in the form that: *to the extent* (whatever extent that may be, partial or complete) that one thinks that self-knowledge is present, *to that extent* only is it justifiable to have the reactive attitudes that define agency." (Bilgrami 2006, p. 116).

[11] Falvey and Owens (1994).

[12] The influence on what follows of the work of Donald Davidson and Robert Brandom (including Davidson (1980) and Brandom (1994), but especially Brandom (forthcoming)) should be obvious.

[13] For a related discussion by Bilgrami himself, see Bilgrami (2006, p. 100).

[14] I'll ignore here the fact that there are some descriptions of our actions that we could be considered *negligent* for being unaware of. (For a discussion of this issue, see Bilgrami (2006, pp. 102 ff.))

[15] Bilgrami (1992a, p. 5).

[16] In Brandom's (1994) terms, the ascriptions that respected a constraint like (C) would be *De Dicto*, while those that did not, would be *De Re*, and as long as we are clear about what

we are doing, there is no reason to deny that either sort of ascription is true. This, of course, requires treating the *de dicto/de re* distinction as one between types of ascriptions rather than as one between types of beliefs, but I think there are independent reasons for doing this, and with the distinction understood this way, Bilgrami's 'Strawsonian' argument will seem much less controversial.

[17] Thanks to Akeel Bilgrami, Gurpreet Raatan, Victoria McGeer, Jack Lyons and audience members at the 2006 meeting of the Canadian Philosophical Association and 2008 meeting of the Southwestern Philosophical Society for comments on earlier versions of this paper.

References

Bilgrami, A. (1992a). *Belief and Meaning*. Cambridge: Blackwell.

Bilgrami, A. (1992b). Can Externalism be Reconciled with Self-Knowledge? *Philosophical Topics* 20(1).

Bilgrami, A. (1999). Why is Self-Knowledge Different from Other Kinds of Knowledge? In Hahn, L. E. (ed), *The Philosophy of Donald Davidson* (pp. 211-224). Chicago: Open Court.

Bilgrami, A. (2006). *Self Knowledge and Resentment*. Cambridge: Harvard University Press.

Brandom, R. (1994). *Making it Explicit*. Cambridge: Harvard University Press.

Brandom, R. (forthcoming). *A Spirit of Trust*. MS, University of Pittsburgh.

Davidson, D. (1980). *Essays on Actions and Events*. Oxford: Oxford University Press.

Falvey, K. and J. Owens. (1994). Externalism, Self-Knowledge, and Skepticism. *The Philosophical Review* 103: 107-37.

Ludlow, P. and N. Martin (eds.). (1998). *Externalism and Self-Knowledge*. Stanford: CLSI Publications.

Nuccetelli, S. (2003). *New Essays on Semantic Externalism and Self Knowledge*. Cambridge: MIT.

Strawson, P.F. (1974). *Freedom and Resentment*. London: Methuen.

Wright, C., B. Smith, and C. Macdonald (eds.). (1998). *Knowing our Own Minds*. New York: Oxford University Press.