Wright on response-dependence and self-knowledge

March 23, 2004

1 Response-dependent and response-independent concepts

1.1 The intuitive distinction

Intuitively, there is a distinction between concepts whose extensions are determined by our responses, and those whose concepts are more or less independent of our responses.

Consider, for examples of concepts which plausibly fit into the former category, *humorous*, *fashionable*, *nauseating*. These are concepts which have determinate extensions - there are some things which fall under them, and some which do not. So there can be ordinary claims with determinate truth-conditions involving these concepts. But, intuitively, the extension of each of these concepts is determined by some response or other of ours. Perhaps whether a joke is humorous depends on whether we would *laugh* at the joke under certain circumstances; perhaps whether something is fashionable depends on whether certain people would either have a certain *belief* or a certain affective response to the item (again, in certain circumstances); perhaps whether or not some substance is nauseating depends on whether it would cause either *feelings* of nausea or the physiological responses associated with them under certain

circumstances. In each case, we have the same pattern: a concept has an extension, but one which is in part determined by our responses (whether these responses are a feeling, a belief, or a physical response). Thus the title 'response-dependent concepts.'

This distinction is interesting in part because it is a plausible way of thinking about the distinction between those properties which are objective, in the sense of being completely independent of us, and those which are not.

Some philosophers, including Wright, have claimed that the class of response-dependent concepts is not limited to these obvious sorts of cases; there are also unobvious, and potentially philosophically interesting, examples. But to see this we will have to have some way, for any given concept, of determining which category the concept falls into.

1.2 Basic equations and the order of determination test

This is the purpose of basic equations, and the so-called 'order of determination test.' Wright's idea (derived from some ideas of Mark Johnston) is this: suppose we want to test whether some concept is response-dependent or not. We then formulate some simple proposition p involving the concept, and put it in a proposition which has the form of a *basic equation*:

$$p \iff \forall S(C \rightarrow S \text{ believes that } p)$$

If we can come up with such an equation for the relevant concept, that is enough to show that there is a tight connection between the extension of the concept and certain of our responses — our beliefs involving the concept in certain specifiable circumstances.

But when will the truth of such a basic equation show that the concept in question is response-dependent? Wright's idea is that such basic equations must satisfy two criteria: (i) it must be possible to state the conditions C in a substantial way which does not presuppose that the concept in question already has a determinate extension, and (ii) the resulting basic equation should be knowable a priori. We will discuss these two requirements in turn.

1.3 The substantiality requirement

The first requirement has two parts: that the conditions C be spelled out in a substantial way, and that they be spelled out in a way which does not presuppose that the concept in question has a determine extension independent of our opinions. It's best to understand these two conditions by working through Wright's discussion of the difference between color concepts and shape concepts.

In the case of color concepts, Wright thinks that we can specify conditions under which one believes that x is red iff x is red: one must observe x in plain view, be attentive, lack doubt about one's perceptual conditions ... (see top p. 79). These conditions are, in Wright's terms, *substantial* because they are not 'whatever it takes' conditions.

Moreover, they do not presuppose that color concepts have an extension independent of our judgements about color. Consider shape concepts, and how one would go about formulating conditions under which one believes that x has a certain (three-dimensional) shape iff x has that shape. We might think that we could specify such conditions for being pear-shaped in much the same way we could specify such conditions for being red. But there is, as Wright notes, a crucial difference. In the case of redness we can spell out 'suitable perceptual conditions' in substantial terms which do not mention redness: in terms of lighting conditions, etc. But in the case of any three-dimensional shape concept, suitable perceptual conditions to ascertaining that some object falls under the concept will involve several perceptions stretching over a period of time (to see the various sides of the object). This by itself need not be problematic; the problem arises from the fact that if the perceptions stretch over a period of time, we must imagine suitable perceptual conditions as ruling out the possibility that the object is *changing shape* during the course of the perception of it. This means that we must build into conditions C the requirement that the object *keep the same shape* during the time it is perceived by the agent in question. But this seems to mean that we are presupposing that there are some facts, independent of our responses, about which shape an object has; and this seems to rule out a response-dependent treatment of concepts of three-dimensional shapes.

1.4 The aprioricity requirement

The second requirement is that the basic equation formed by specifying, for the concept to be tested, substantial conditions under which beliefs covary with the extension of the concept, should be knowable a priori. There are two (related) ways to see why this might intuitively be a good test for response-dependence.

First, note that the right hand side of basic equations (which satisfy the substantiality condition) are logically independent of the left-hand side. So we need some explanation of how such equations could nonetheless be knowable a priori; the implicit idea is that such equations will be conceptual truths, which anyone competent with the concept could know. But if this is so, then it seems that the extension of the concept is not independent of our 'best opinions' (i.e., our beliefs in conditions C).

Second, suppose that the basic equation were a posteriori. Then it should be conceivable that it is false: i.e., it should be conceivable that either something falls under the concept even though we could not ascertain that it does, or that our best judgement would be that something falls under the concept even though it does not. But to imagine this sort of thing happening just is to imagine the extension of the concept being fixed independently of us.

2 Wright on the response-dependence of concepts of mental states

Wright takes Wittgenstein to have shown that there can be no explanation of our capacity for self-knowledge if we conceive of facts about our mental states and their contents as purely objective: as independent of our opinions about those mental states. He then thinks that he can supply a view of those mental states which can explain self-knowledge.

2.1 Wright on Wittgenstein on rule-following

Wright takes Wittgenstein's target in the famous 'rule-following' sections of the *Philosophical Investigations* to be the view that facts about the contents of our mental states are response-independent.

A first step in getting to this argument is to note that properties like meaning something by an utterance or having a belief or intention are located between two extremes: 'conscious' states, like perceptual or bodily sensations, and character traits, like honesty or bravery. They are like the second in the sense that they seem to be constitutively tied to an agent's dispositions; but they are like the first in that we have a privileged kind of self-knowledge of them. The problem in giving an adequate accounts of such states, then, is to explain our privileged access to them without, falsely, assimilating them to the case of sensations.

So the first argument against a kind of introspectionist epistemology of self-knowledge is that there is nothing that it is like to have a belief, or an intention, or to mean something, and so nothing to introspect. (As we have seen, Peacocke would disagree here, and claim that knowledge of our own beliefs is based on conscious judgings; it is not obvious how this would apply to knowledge of meaning.)

But there is a further problem, and this is where Wittgenstein's arguments supposedly come in. The background to this discussion is that the meaning of an expression determines its applications in future cases; so, e.g., the fact that by '+' we mean addition determines the fact that the sentence '67+58=125' is true. Further, the idea is that when we know the meaning of an expression, we know how to go on to apply it to further cases (though of course not infallibly). This generates a problem for the introspection-theorist: how can awareness of some kind of 'meaning-sensation' tell me how to go on in future cases, as it must if that meaning-sensation is to constitute my knowledge of the meaning of a word? (This is different than the first sort of argument in that it assumes for reductio that there is such a thing as a meaning-sensation, and argues that it cannot do the work that it is meant to do.)

What we are supposed to recognize is that the connection between a meaning-sensation and applications of a term to other cases is not an automatic one. Once one realizes this, the natural response is that the meaning-sensation would, itself, require some sort of interpretation. One form this might take is a verbal expression of a rule associated with the meaning-sensation. But, as Wittgenstein asks in §198 of the *Investigations*,

> "'. . . how can a rule show me what I have to do at *this* point? Whatever I do is, on some interpretation, in accord with the rule.' - That is not what we ought to say, but rather: any interpretation still hangs in the air along with what it interprets, and cannot give it any support. Interpretations by themselves do not determine meaning."

The idea here is that if we are wondering what knowledge of meaning can consist in, any appeal to special meaning-sensations will fail, because such sensations will have no connection to future uses of the term, whereas knowledge of meaning does. Neither will it help to suppose that we associate each meaning-sensation with some formulation of a rule for use, because then we will just be appealing to the kind of fact we were trying to explain: knowledge of meaning. (We will have to know the meaning of the verbal formulation of the rule; but this was supposed to explain knowledge of meaning in general.)

This indicates, as Wright sometimes puts it, that there can be no substantial epistemology of rule-following. But we need some explanation of this fact; knowledge of meaning is not magic. Wright's view seems to be that so long as we view the concept of meaning as response-independent – and so view facts about what we mean as existing independently of our best judgements about those facts – we will simply have no solution to this mystery.

(One question worth asking at this point: how does this point cut against a Shoemaker-style reliabilist theory, which claims that there is a kind of substantial epistemological story to be told about self-knowledge, but crucially not one which can be told in terms of what is available to introspection?)

One question about this argument is: can it be made to apply to the case of our knowledge of our own mental states, as well as our knowledge of what we mean by our words? The obvious application seems to be this: just as knowledge of the meaning of a word involves knowing how to apply it to new cases, so knowing the content of a belief involves knowing the conditions which would make that belief true, and so knowing the content of an intention involves knowing what actions would satisfy the intention. But just as a meaning-sensation does not involve one particular way of going on to apply a term to new cases, so a belief-sensation does not automatically lead to a grasp of conditions which would make the belief true, and so on . . .

Wright's solution to the problem posed by the rule-following argument is that concepts of the states of which we have privileged self-knowledge are response-dependent. This breaks down into two parts: the argument that concepts of mental states are response-dependent, and the explanation of how this is supposed to help explain self-knowledge.

2.2 Are mental concepts response-dependent?

We know now that mental state concepts will be response-dependent if they pass the order-of-determination test; if, that is, we can formulate a basic equation which is

both substantial and knowable a priori.

Consider the case of intention. Wright seems to think that the following is knowable a priori:

> S intends to $\phi \iff \forall S((S$ is attentive & S grasps the concept of intention & S is not self-deceived$) \rightarrow (S$ believes that he intends to $\phi))$

This seems plausible; but, as Wright notes, the conditions built into this basic equation are not substantial; as it stands, ruling out self-deception is just a whatever-it-takes condition of the illegitimate kind.

One response to this would be to try to analyze self-deception in substantial, not intention-presupposing terms. Wright does not attempt this.

Instead, he asks us to focus on the basic equation obtained by deleting the self-deception condition:

> S intends to $\phi \iff \forall S((S$ is attentive & S grasps the concept of intention$) \rightarrow (S$ believes that he intends to $\phi))$

Some instances of this basic equation will be false, because there are some cases of self-deception. But Wright claims that if we take any one instance of this equation, say

> Jones intends to $\phi \iff$ ((Jones is attentive & Jones grasps the concept of intention) $\rightarrow$ (Jones believes that he intends to ϕ))

we will be a priori justified in believing it. So we have given a basic equation for intention which is not a priori knowable, nor even a priori credible, but rather has instances all of which are a priori credible. (Presumably a proposition is a priori credible iff one is justified a priori in believing it.)

This is a substantially weaker claim. But Wright thinks that we still owe some explanation of the fact that these instances are a priori credible; if having a certain intention really were completely conceptually independent of believing of oneself that one has that intention, then why wouldn't the right default attitude toward instances of this basic equation be one of suspension of belief?

2.3 Response-dependence and self-knowledge

Wright's explanation, of course, is that the default position is assenting to instances of this basic equation because the concept of intention is response-dependent: the extension of the concept of intention is partly determined by one's views about one's own intentions. As Wright puts it:

> "... there are non-trivially, independently specifiable conditions whose satisfaction ensures, courtesy of no *a posteriori* background beliefs, that, failing any other relevant information, a subject's opinions about his or her intentions should be accepted. And the proposed strategy of explanation is ... [that] what determines the distribution of truth-values among ascriptions of intention to a subject who has the conceptual resources to understand those ascriptions and is attentive to them are, in the first instance, nothing but the details of the subject's self-conception in relevant respects." (252-253)

This is not meant to rule out the possibility of error; sufficient discordance between behavior and stated self-conception may be enough in some cases to count ascriptions of intentions to oneself as false.

A couple of questions about this account:

1. Does it carry over easily to belief, meaning, and the other cases in which one might be interested?

2. How should we think of animal intentions in the light of this account?

3. Does the possibility of self-deception in the end show that there must be facts about intentions which are independent of the subject's self-conception?

3 Objections

We will consider a number of objections to Wright's proposal. The first and third are from Appendices I and III to Mark Johnston's 'Objectivity refigured', and the second is from Paul Boghossian's article on rule-following.

3.1 Johnston's missing explanation argument

In Appendix I to 'Objectivity refigured', Johnston presents a very simple argument against the possibility of Wright-style proofs of the response-dependent nature of concepts. Wright's arguments depend on certain substantial basic equations being knowable a priori. Johnston's argument is that no philosophical interesting substantial basic equations can be true because of the following principle:

> If it is true that q because p, and p is a priori equivalent to $p*$, then the claim that q because $p*$ must be true as well.

But the two propositions which Wright claims to be a priori equivalents in the case of basic equations do not satisfy this principle. Consider, e.g., the concept of being red. The proponent of a response-dependent treatment of redness says that

> x is red

is a priori equivalent to

> $\forall S$ ((S has x in plain view, is attentive, lack doubt about his perceptual conditions ...) $\rightarrow$ S believes that x is red)

But this can't be right, since

> John believes that x is red because x is red.

is true, whereas

> John believes that x is red because $\forall S$ ((S has x in plain view, is attentive, lack doubt about his perceptual conditions ...) $\rightarrow$ S believes that x is red).

or

> John believes that x is red because John has x in plain view, is attentive, lack doubts about his perceptual conditions ... & John believes that x is red.

are not. The former claim is a genuine causal explanation which claims that John has a certain belief because of his contact with the state of affairs of x being red, whereas the latter two are, as Johnston says, 'explanatory solecisms.' We cannot read the latter statements as giving a *cause* of John's having this belief.

The same argument applies, it seems, to any Wright-style response-dependent analysis of mental concepts. Suppose that I believe, corrrectly, that I intend to come to McGill tomorrow. Then I can truly say,

> I believe that I intend to come to McGill tomorrow because I intend to come to McGill tomorrow.

whereas I am not explaining my belief by saying

> I believe that I intend to come to McGill tomorrow because I am attentive, am not self-deceived, and believe that I intend to come to McGill tomorrow.

The only concepts for which response-dependence is plausible will be those concepts for which the corresponding explanations cannot be formulated; plausibly, the claim that S is nauseated by x because x is nauseating does not count as a genuine causal explanation. If this is right, then the concept of the nauseating seems amenable to response-dependent analysis. But no interesting philosophical cases seem to be of this kind.

This argument is convincing if you buy the premise that a priori equivalents should be substitutable in (causal) explanations. This premise seems plausible; is there any reason to doubt it?

3.2 Boghossian's claim of incoherence

About Wright's view of the response-dependent nature of mental states, Boghossian argues:

> "In a way, an intuitive difficulty should have been clear from the start. A 'judgement-dependent' conception of a given fact is, by definition, a conception of that fact according to which it is *constituted* by our *judgements*. This idea is clearly appropriate in connection with facts about the *chic* or the *fashionable*; familar, though less clearly appropriate, in connection with facts about colour or sound; and, it would appear, impossible as a conception of facts about mental content. For it cannot in general be true that facts about content are constituted by our judgements about content: facts about content, constituted independently of the judgements, are presupposed by the model itself." (547)

Recall Wright's discussion of what it takes for a basic equation to be substantial: the conditions specified must not presuppose that the concept in question has an extension, independently of our judgements involving it. We found that shape concepts were response-independent, on the grounds that the conditions under which beliefs and extension co-vary involve the object in question maintaining the same shape — and this seems to presuppose that there is a coherent notion of an object maintaining a certain shape independent of our judgements.

Boghossian's point is that the very form of basic equations shows that we will always reach this verdict with respect to our beliefs. For a basic equation for any response-dependent concept will analyze that concept in terms of some set of favorable conditions obtaining *and* the agent in question having a certain belief. In particular, if we were to give a response-dependent analysis of the concept of belief, the analysis would include certain facts about our beliefs. Doesn't this show that the analysis presupposes the existence of certain facts about belief independent of our best opinions, just as the corresponding result in the case of shape shows the response-independence of shape?

Boghossian seems to think that this applies to all facts about 'mental content' — and this includes not just beliefs, but also intentions, desires, and all other contentful

mental states. There seems to be no argument for expanding this result about belief to all mental states; but it would be odd if belief, lone among our mental state concepts, were response-dependent.

3.3 Johnston on the formulation of Wright's criterion

Johnston presents a more general argument against Wright, in the form of a dilemma. He writes:

> "Wright's distinction turns on the directional idea of determination. But how can the extension of a proposition P be determined under specified circumstances by anyone's belief that P? Surely only if the belief that P already has some extension associated with it. For if the belief that P has no extension and so no mode of determining an extension associated with it then the belief that P will be devoid of content and so will not constrain anything at all."

This seems to show that the idea that a belief (or pattern or beliefs) involving a concept should fix the extension of the concept is incoherent: the idea that a belief involves the concept already involves the idea that the concept is contentful, and so that it has a relation to a reference.

Johnston concludes that there are two ways for Wright to go. Either (i) the responses in terms of which basic equations for concepts are given cannot involve the concept (whether or not the concept occurs in the content of an ascribed belief), or (ii) we should give up the idea that there is a one-way dependence here, and settle instead for the idea that there can be response-interdependence between some domain of facts and our best opinions about those facts.

(i) does not seem promising, since reductive analyses of concepts in terms of our responses (where those responses do not involve any relation to the concept in question) are implausible.

(ii) is more plausible, but does not reveal anything about whether the extension of a concept is determined by facts about our responses.

More to the point for us, interdependence theories do not obviously promise any kind of explanation of privileged access to facts about our own mental states.