Intentions, Response-Dependence, and Immunity from Error[1]

RICHARD HOLTON, MONASH UNIVERSITY

You are, I suspect, exceedingly good at knowing what you intend to do. In saying this I pay you no special compliment. Knowing what one intends is the normal state to be in. And this cries out for some explanation. How is it that we are so authoritative about our own intentions? There are two different approaches that one can take in answering this question. The first credits us with special *perceptual* powers which we use when we examine our own minds. On this view we *detect* our own mental states in much the same way that we detect the state of the world around us; but the powers we direct inward are much less prone to error than those we direct outwards. The alternative approach denies that there is such a thing as inward perception. On this view the whole idea the we detect our own mental states using some kind of internal perceptual apparatus is misguided; a wholly different account is needed.

Crispin Wright has embraced the second of these approaches. In an extremely stimulating article he has proposed that our concept of intention is *response-dependent*.[2] (Or at least he has proposed that it is an *extension-determining* concept, and as we shall see that's pretty much the same thing.) This he thinks provides an explanation of why we are so good at knowing what we intend; and it does so, he claims, by subverting the idea that we *detect* our own intentions.

Wright's article will be the focus of this paper; but in order to discuss it properly I have to do a fair amount of preliminary work. Lest the trees hide the wood, let me spell out my strategy in some detail. In the first section I introduce the notion of a response-dependent concept. In the second I discuss the relationship between response-dependence and immunity to error. Being immune to error is the limit case of being good at knowing. If we treat intention as a response-dependent concept, will this entail that we cannot be in error about our intentions? A suggestion of Philip Pettit's entails that they would: he has suggested that *all* response-dependent concepts confer immunity from error upon those who use them. But I argue that Pettit's suggestion is not correct. Only some response-dependent concepts confer immunity from error on those who use them; I give an account of which ones do.

In the third section I turn to Wright, and argue that for my purposes there is no significant difference between his notion of an extension-determining concept, and the notion of a response-dependent concept. This gives me licence to apply to him the lessons learned in discussing Pettit. That completes the preliminaries. In the fourth section I start on the main topic: I discuss Wright's argument that intention is a response-dependent concept. It's a complex

---

[1]Versions of this paper were presented at Princeton and Monash, and at the Workshop on Response Dependence at the ANU in June 1991. I am grateful for the comments that were made on these occasions. Special thanks are due to Mark Johnston, Rae Langton, Philip Pettit, and Michael Smith.

[2]Crispin Wright, 'Wittgenstein's Rule-following Considerations and the Central Project of Theoretical Linguistics' in A. George (ed.) *Reflections on Chomsky* (Oxford: Basil Blackwell, 1989) pp. 233-64. See especially Sections III and IV.

argument. Wright presents it in just five pages, but I spend considerably more in evaluating it. His claim is not that we are immune to error in assessing our own intentions, but rather that there is a presumption that our assessments will be correct. From this premise he concludes that the concept of intention is response-dependent. I present two ways of understanding his argument; on both it fails. Nevertheless, I think that there is something to be salvaged from it. This is the contention of the fifth section. There I claim that a certain prominent theory of mind, namely analytic functionalism, does treat the concept of intention as response-dependent; and, on certain reasonable assumptions, it entails the same kind of presumption in favour of the accuracy of agents' assessments of their own intentions that Wright claims to hold. This conclusion represents something of a vindication of Wright's position; but only a partial vindication. Wright's contention, recall, is that to embrace a response-dependent account of intention involves rejecting a perceptual epistemology of mind; it involves giving up on the idea that we *detect* our own intentions. This, I think, is simply wrong. In the sixth section I explain why.

I RESPONSE-DEPENDENT CONCEPTS

What are response-dependent concepts? The intuitive idea is that they are concepts whose extensions are essentially determined by human responses. Some concepts are clearly response-dependent. Take the concept of being *irritating*. Irritating things are things that provoke a particular response - irritation - in the people who are unfortunate enough to come into contact with them. Similarly, nauseous things are things that provoke nausea. Exciting things are things that provoke excitement. Tiring things are things that provoke fatigue. Soporific things are things that provoke sleep. Tedious things are things that provoke something, but it's not so easy to say what it is. Perhaps it is the judgement that they are tedious.

Besides the concepts which are obviously response-dependent come those concepts which are arguably so. The theory of secondary qualities may be seen as an attempt to cast the net wider. So colour, sound, taste, and warmth have all been treated as response-dependent concepts.

So far, so good. But we have a problem. We are interested in those concepts which are *essentially* connected to certain responses. We must find a way of distinguishing them from concepts which are typically connected to certain kinds of response as a matter of accidental fact. For instance, it might be the case that square things will, in fact, typically provoke the judgement that they are square. But we don't want to class the concept of square as response-dependent: the connection between the concept and the response is somehow too accidental. How can we make the idea precise? An obvious way is to say that there is an *a priori connection* between, say, something being red and it being judged to be red; and that there is no such connection between something being square and it being judged to be square. We can make the idea more precise still. Suppose we construct a biconditional equation. On the left-hand side we put the concept in question; on the right, a response to which we suspect the concept is tied. Following Mark Johnston, call this a *basic equation*.[3] Now we can say that a

[3]For a discussion of such equations see his articles 'Dispositional Theories of Value' *Proceedings of the Aristotelian Society*, Supplementary Volume 63 (1989)

concept is a response-dependent concept if and only if at least one such equation is *a priori true*. The *a priori* truth of the equation reflects the conceptual connection between the concept and the response. If red is a response-dependent concept this will be shown by the fact that something like:

> x is red iff x would be judged to be red under conditions C

is *a priori true*. Whereas if square is not a response-dependent concept this will be shown by the fact that a similar equation for square:

> x is square iff x would be judged to be square under conditions C

is, if true, at most *a posteriori true*.[4]

Two qualifications. First, the basic equation for red will only be *a priori* true if the judgements which it invokes are those of normal observers in standard conditions. Colour blind people won't always judge red things as red, nor will those in pitch black rooms. This caveat is covered by the C-conditions. They specify what normal observers and standard conditions are. For instance, they might restrict the observers to those who are statistically normal, who have the concept of red, and who are paying attention; this would rule out the colour blind, the very young, and the unconscious or otherwise inattentive. And they might limit the the circumstances of observation to those which do not involve strange lighting. We must, however, be careful about what we put in the C-conditions. If the basic equation is to be *a priori* true, we must in turn have *a priori* knowledge about what the C-conditions are.

Second qualification. I have suggested that basic equations can be used as a device for determining which concepts are response-dependent. They will be those which have *a priori* true basic equations. But this is not quite right as it stands. It is possible to construct *a priori* true basic equations for any concept if we allow what Wright calls 'whatever-it-takes' formulations of the C-conditions, i.e. formulations that make the basic equation *a priori* true in a trivial way. As an example of a 'whatever-it-takes' formulation, suppose we specified the C-conditions on the basic equation for red by saying that the only circumstances to be considered are those in which observers are accurate in their perceptions of what is and is not red: then the biconditional would clearly be *a priori* true. But this shows us nothing about the nature of the concept red, since we can construct a similar *a priori* basic equation for just about any concept. For instance, we can construct one for square:

> x is square iff x would be judged to be square by observers who are accurate at identifying square things in circumstances which are propitious for doing so

---

pp. 139-74; and 'Objectivity Refigured: Pragmatism Without Verificationism' in John Haldane and Crispin Wright (eds.) *Realism and Reason* (Oxford: Oxford University Press, forthcoming).

[4]I have given a basic equation for red that exploits the idea that red things *provoke judgements* that they are red (a judgement equation), rather than one exploiting the more familiar idea that they simply *appear* red (an appearance equation). I do so because of the special implication of the judgement equation for the possibility of error: see below, Section II. Of course, someone could reject the idea that red has an a priori true judgement equation, and yet maintain that red is response dependent on the grounds that it has an a priori true appearance equation.

This equation is *a priori* true, but that tells us nothing about the concept of square, since it is trivially true: the normal observers are specified as those who are good at perceiving square things, the standard conditions as those that are propitious for so doing. So, if we are to use the basic equation to tell us anything interesting about the concept it contains, we had better ensure that the C-conditions are not formulated in a 'whatever-it-takes' way.

Can we formulate C-conditions for red which, one the one hand, we know *a priori* to be sufficient to rule out abnormal observers and non-standard conditions, and yet which, on the other, do not trivialize the basic equation? That's an open question, but I won't inquire into its truth here. Let's assume that it can be done.

Filling out the idea: preliminary points about response-dependence

I'll make five points about the notion of response-dependence. The last of these will be relevant later; but my main aim is just to ensure that the notion is clear. First, note that response-dependence is a feature of concepts, not of properties. Suppose there were a straightforward characterization of the reflectance property that an object must have to be seen as red, a characterization which is given by a description couched in the language of physical science. The concept expressed by that description picks out the same property as is picked out by the concept of red; but it is not a response-dependent concept. There is no *a priori* connection between it and a response.

The second point to note is that whilst the basic equation does, in a sense, provide an analysis of the concept of red, it does not provide a *reductive* analysis. It does not analyse red using other concepts in such a way that we could eliminate talk of red altogether. This is clearly so, for the concept of red occurs on both sides of the biconditional. Third, the basic equation is claimed to be a priori; it is not claimed to be necessary. Treating red as a response-dependent concept is quite compatible with Kripke's reference fixing account, according to which the basic equation is a contingent *a priori* truth.[5] Fourth, note that on the basis of the apriority of the basic equation alone we cannot infer much about the world. We cannot conclude that there are red things, because it is an empirical matter whether there are things that appear red. Even the truth of the sentence 'red things appear red' will depend on whether or not there are things which appear red, and that is not an *a priori* matter.[6] Equally, there will be no *a priori* inference from the fact that something has been judged red to the conclusion that it is red, because it will be an empirical matter whether or not the C-conditions were fulfilled on that occasion.

Finally, note that not just any colour concept is response-dependent. Suppose I define the colour word 'extravagance' as denoting the predominant exterior colour of Prince Charles' Bentley.[7] I'm confident that extravagance is a colour with which I am acquainted. Bentleys are exclusive cars, but not so

---

[5]Saul Kripke, *Naming and Necessity*, (Cambridge: Harvard University Press, 1980) p. 140 n. 71.

[6]Or rather it will if we treat 'red things' as a definite description ('the things which are red') and then analyse that on standard Russellian lines

[7]Perhaps Prince Charles has more than one Bentley. I mean the one that was driven from England to Czechoslovakia for his official visit in May 1991 consuming petrol at fifteen miles to the gallon.

exclusive as to demand entirely new colours. The concept of extravagance is a colour concept, but it is not response-dependent. I am a normal perceiver of colours, but if I were confronted with a sample of extravagance I wouldn't identify it as such. I wouldn't because I have no idea what colour extravagance is; nor, I suspect, do any other users of the concept. So some colour concepts are not response-dependent. Which ones are? We could, of course, just list them, but it would be good to say what those we would list have in common. For a first approximation we might say that a colour concept will be response-dependent if and only if the colour it denotes is one with which the users of the concept are acquainted under that very concept; but I don't know how to make the notion of 'acquaintance under a concept' precise.[8]

II IMMUNITY FROM ERROR

Part of the interest in response-dependent concepts has been provoked by the thought that they confer a form of immunity from error. And we have seen that a somewhat weaker claim, the claim that we are very good at identifying our own intentions, has provided a motivation for thinking that self ascriptions of intention are response-dependent. In this section I'll investigate just how response-dependence and immunity from error are related.

It appears that *some* response-dependent concepts do indeed confer immunity from error. Suppose that it is *a priori* true that something is red if and only if normal observers would judge it to be red in standard conditions. It follows that what normal observers do in fact judge to be red in standard conditions will be red. Normal observers will not make errors of commission: they will not judge things to be red when they are not. And they will not make errors of omission: they will not fail to judge things as red when they are. It follows that, by and large, we will not be wrong about what is red (quite what we mean by 'by-and-large' depends on how we define normal observers and standard conditions). But how far does this immunity go? Philip Pettit thinks that it extends to response-dependent concepts quite generally. He writes:

> [A]s an observer under normal conditions cannot be in ignorance or error about the colour of something, so the responses involved in *any* response-dependent area of discourse cannot lead subjects astray under those conditions.[9] (My italics)

This is a striking claim. If it is true then error will be ruled out in very many areas. But is it true? Surely we could be completely wrong about which things are soporific. We could fail to identify what it is that is putting us to sleep. Similarly we could be wrong about which are nauseous: we could be wrong about

---

[8]I think that this is roughly right for the *hues*; but once we are acquainted with them it appears that we do not need to be independently acquainted with each *shade* of each hue. It seems reasonable to think that if the concept of blue is response dependent then so is the concept of Hume's missing shade of blue, even if no one is acquainted with that shade.

[9]Philip Pettit, 'Realism and Response-dependence' *Mind* 100 (Oct 1991), and reprinted in this collection. Pettit only claims that immunity from error is guaranteed under ideal conditions of enquiry. My criticisms apply even there - provided, of course, that the ideal conditions are not defined in a whatever-it-takes way as those in which there will be no error.

which things are causing us to feel nausea.[10] And we could be wrong about which things are sexually arousing for a different reason: we can, it appears, be wrong about when we are aroused.[11] We need to say what is special about red: what it is about it that brings immunity from error. And to do this we need to make some distinctions.

Varieties of response-dependence

There are three important features of the concept of red which are not shared by all response-dependent concepts. Firstly, the responses which figure on the right-hand side of the basic equation for red are judgements. I'll call a concept which is response-dependent in this way a *judgement-dependent* concept. Secondly, the judgments that figure in the basic equation are judgements about what is red; that is, they are judgements about the applicability of the very concept which is being analysed. I'll call concepts which have basic equations of this form *echo* concepts: the concept that figures on the left-hand side is echoed on the right.[12] Thirdly, the class of people whose judgements figure in the basic equation for red - the judges - is pretty much coextensive with the class of people who use the concept of red. I'll call a concept for which this is true a *users'* concept.

In these respects red provides a paradigm to which other response-dependent concepts approximate. Some response-dependent concepts are clearly not judgement-dependent: for instance, let's stipulate a technical use of *stimulant*, according to which a stimulant is something that causes people's hearts to beat faster. It is *a priori* true that something is a stimulant if and only if people would typically respond to it with an increased heartbeat. The response in question is not a judgement, so the concept of a stimulant is response-dependent without being judgement-dependent. Other concepts stand on the border between the judgement-dependent and the merely response-dependent. Is something funny because it induces laughter or because it induces the judgement that it is funny? Is a sentence grammatical because competent speakers accept it, or because they judge that it is grammatical?[13]

Amongst the judgement-dependent concepts some are echo concepts and some are not. Suppose an atheistic anthropologist introduces the concept of *religious significance*: the anthropologist stipulates that an object is religiously significant if and only if the faithful judge it to be sacred. The concept of

---

[10]Note carefully the kind of error it is that I say is possible. Many have thought that if someone is feeling nauseated then they must know that they are. I do not deny it. What I deny is that someone who is caused to feel nausea by an object is guaranteed to know which object it is that is causing their nausea. And if they do not know which object it is that is causing their nausea they will not know which object it is that is nauseous.

[11]On people's inability to tell whether or not they are sexually aroused, see Edward Donnerstein, Daniel Linz and Steven Penrod, *The Question of Pornography* (New York: The Free Press, 1987).

[12]It is this feature which entails that the basic equation for red cannot be a reductive definition of red.

[13]Johnston defines the response dependent more tightly than me by restricting it to those concepts that have basic equations involving a mental response. See 'Objectivity Refigured'. The result of this is not, however, to exclude all the non-judgement dependent concepts. I can have a mental response that is not a judgement.

religious significance is judgement-dependent, but it is not an echo concept: the concept on the left of the basic equation is not the same as the concept on the right. To be religiously significant is not to be sacred, as can be seen from the fact that the atheist anthropologist believes that there are no sacred objects, but that there are religiously significant ones.

Religious significance also provides an example of a concept which is not a users' concept: the users of the concept, the anthropologist and his ilk, are not the same group of people as the judges of the concept, the faithful. But these two features - being a echo concept and being a users' concept - do not always go hand in hand. Some concepts are echo concepts but not users' concepts. Consider an example discussed by Pettit: the concept of U.[14] Something is U if and only if a certain group of people, the Sloanes, would judge it to be U. U is thus an echo concept. The Sloanes comprise a very small part of the population; but the concept of U, let us accept, is regrettably widely used. So the class of users of the concept is not coextensive with the class of judges of the concept: U-ness is not a users' concept.[15] Another example: there could be games (perhaps there are) in which a team has a certain score if and only if the referee judges that it has that score. The score is of great interest to many beside the referee: players, managers, spectators, bookmakers. They all make use of the concept of the game's score, but their judgements do not figure in the basic equation; so it is a echo concept but not a users' concept. Equally there can be users' concepts which are not echo concepts. Let us say, crudely, that something is popular if and only if most people judge that they like it. The class of users is roughly coextensive with the class of judges, so popularity is a users' concept. But it is not a echo concept: the judgements in question are not those about whether the thing is popular but about whether it is liked.

Enough new jargon. Let me recap. A response-dependent concept will have a basic equation that makes reference to a response of some kind. If the response is a judgement, I call the concept a judgement-dependent concept. If it is a judgement about the applicability of the very concept that is being analysed, I call the concept an echo concept. If the class of people whose responses figure in the equation is roughly coextensive with the class of people who use the concept, I call it a users' concept.

We are now in a position to return to the question of immunity from error. I suggest that the only response-dependent concepts which automatically confer immunity from error on those who use them are *judgement-dependent concepts which are both echo concepts and users' concepts.*[16] Other response-dependent concepts confer no such guarantee.

Consider first response-dependent concepts that are not judgement-dependent. As an example, recall our technical usage according to which something is a stimulant if and only if it causes people's hearts to beat faster. We could make both errors of commission and errors of omission in applying the concept of being a stimulant: that is, we might classify some things as stimulants

---

[14]'Realism and Response-dependence'

[15]The fact that the Sloanes are both users and judges of the concept is not enough to make it a users' concept.

[16]I say 'automatically' because there might be other response dependent concepts which confer immunity from error, but which do so because of some further feature they possess which is independent of their response dependent nature.

when they are not, and we might fail to classify other things as stimulants when they are. First, the errors of commission. We could misidentify the relevant *responses*: we could think that our hearts were beating faster when they were not.[17] Or we could misidentify the *cause* of the relevant responses: we could think it was the chilli, when in fact it was the tea. Second, the errors of omission. These could be caused by any of the factors that lead to errors of commission. But in addition, since the response itself is independent of any judgement about that response, we might simply fail to take the further step of making the judgement. We might specify the C-conditions in such a way that only those who are paying attention will count as normal observers. But this will not guarantee that they will go on to make the relevant judgements. We might notice that our hearts are beating faster, but, having noticed, we might fail to conclude that the thing that caused it is a stimulant.

Similarly, we might be quite generally in error in applying judgement-dependent concepts if they are not both echo concepts and users' concepts. Consider the case of popularity. According to our crude definition, something is popular if and only if most people judge that they like it. Popularity is a judgement-dependent concept; and since the judges are the users, it is also a users' concept. But it is not an echo concept: the concept of being popular is not the same as that of being liked. Are we immune to error in our judgements about what is popular? No. We might all be wrong about what is popular, since we might be wrong about what other people like. Each of us secretly hates Big Brother, and yet, thinking the others don't, each concludes that he is popular.

The case of response-dependent concepts which are not users' concepts is more complicated. Suppose we have a concept which is used by the people whose responses feature in the basic equation, but which is also used by very many people whose responses do not feature there. If the concept is a judgement-dependent echo concept, like U, we will have a guarantee that the judges of the concept are, by and large, immune to error; but this guarantee will not apply to those who use the concept yet are not judges of it. The majority of people, not being Sloanes, could be quite wrong about what is U and what is not. When we turn to a concept like religious significance for which *none* of the judges are users, we find that none of the users of the concept will be immune to error. The anthropologists could be quite wrong about what the faithful find sacred.

So what shall we make of Pettit's claim that in applying response-dependent concepts we will in general be immune to error? Pettit argues as follows: we are immune to error in our judgements about what is red; all response-dependent concepts are like red in the relevant respects; so we are quite generally immune to error in our judgements involving response-dependent concepts. We have seen that this is not a good argument. The second premise is false. All response-dependent concept are not like red in the relevant respects. Red is a judgement-dependent concept which is both an echo concept and a users' concept; this is what brings the immunity from error. But not all response-dependent concepts are like that. If we seek to explain why it is that we make so few errors in

[17]Of course, we might have grounds for thinking that there are some responses we could not misidentify: that people could not misidentify their own pain for instance. But these are *independent* grounds for ascribing immunity to error. The immunity does not accrue to the concept simply in virtue of the fact that it is response dependent.

identifying our intentions on the grounds that the concept of intention is response-dependent, we will need to take care in determining just what kind of response-dependent concept it is.

III EXTENSION-DETERMINING CONCEPTS

We can now turn to Wright's paper. Wright argues that intention is an extension-determining concept. This is tantamount, I say, to arguing that it is a response-dependent concept. In this section I will say why.

According to Wright, extension-determining concepts are concepts which have their extension fixed by our best judgements: by our best judgements as to which things satisfy those concepts. Suppose F is an extension-determining concept. Then there is a conceptual link between something being F and our judging it to be F. So we might think that the extension-determining concepts are those which we have called judgement-dependent echo concepts.[18] I think that this is the right conclusion. And it is a conclusion which appears to be born out by the fact that the criterion Wright uses to tell which concepts are extension-determining is one which makes use of the *a priori* status of certain biconditionals. Nevertheless, it is not the same criterion that we have been using. I need to say why this is.

We have used the basic equation as our criterion for response-dependence, and a particular form of that equation as a criterion for judgement-dependence. We have said that red will be shown to be a judgement-dependent echo concept if the equation

(1) x is red iff (x would be judged to be red under conditions C

is *a priori* true. Wright proposes an proposes an different criterion for identifying the extension-determining concepts. He says that red will be shown to be an extension-determining concept if the following equation (which he variously terms the *provisional* or *provisoed* equation) is *a priori* true:

(2) If the C-conditions hold, (x is red iff x is judged to be red)

Why does Wright propose a different criterion? The reason is that he thinks there is a problem with using the basic equation. The problem concerns the possibility that the situations in which the C-conditions obtain are situations which will distort the colour of the object we are considering. Mark Johnston has provided an example. He asks us to imagine a shy but intuitive chameleon. At the moment it sits in the dark and it is green. But were we to reach for the light, it would know, and would blush red. This will present a problem if we construe the basic equation as

[18]Whether they are users' concepts will depend on who uses the concept, and whose judgements get to count. We will return to this issue shortly.

(1*) x is red iff (if the C-conditions were to obtain, then x would be judged to be red).

The problem is that were the C-conditions to hold, were there enough light to see the chameleon, the chameleon would be red. And on a standard counter-factual reading of the basic equation, this will entail that *the chameleon is red even as it sits in the dark*: it is true of the chameleon as it sits in the dark that if the C-conditions were to hold, it would be judged to be red, since if the C-conditions were to hold it would be red. But that is just the wrong result. The chameleon is green as it sits in the dark.

How should we react to this problem? One possibility is to give the *if-then* clause that occurs in the basic equation a non-standard counterfactual reading: we might propose that the possible world with respect to which the judgement of the observers are to be evaluated is not the world which is closest to the actual one, but instead is the closest world in which the chameleon is not both shy and intuitive. A second possibility is to reject (1*) as the correct construal of (1). On such an approach where (1) says that x *would* be judged to be red, this is understood this to mean that x *is disposed* to be judged as red, where this disposition cannot be straightfowardly analysed in terms of a counter-factual conditional. This is the approach that Johnston takes.[19]

Wright takes a different approach. He proposes that we should reject the basic equation as a criterion. In its place he proposes the *provisional* equation, i.e.

(2) If the C-conditions hold, (x is red iff x is judged to be red)

A concept is extension-determining according to Wright's account if and only if a relevant provisional equation is *a priori* true. What is the idea here? It is to restrict the applicability of the criterion to circumstances in which the C-conditions obtain. So the case of the shy but intuitive chameleon which sits in the dark no longer presents a problem. The C-conditions do not obtain there, so the criterion cannot be applied.[20]

It's debatable whether this response to the problem of the shy but intuitive chameleon is superior to the responses outlined above.[21] But this won't be my concern here. The point I want to stress is that Wright is proposing that we

---

[19]'Objectivity Refigured'

[20]'Wittgenstein's Rule-following Considerations' p. 261 n. 25. See also his 'Moral values, projection and secondary qualities', *Proceedings of the Aristotelian Society*, Supplementary Volume 69 (1988), n. 26. The fact that the criterion cannot be aplied does not entail that the chameleon has no colour as it sits in the dark. It just entails that we cannot use the provisional equation to determine which colour it has.

[21]In particular, Johnston has raised various other difficulties which would pose problems for Wright's account, but which might be handled using a dispositional account of the kind he favours. Imagine, for instance, another chameleon, this time with a translucent green skin. As a result of the workings of its digestive system, the chameleon radiates a brilliant red light from its belly, which causes all who see it to judge its skin as red. We might hope for a dispositional account according to which the skin *is disposed to be judged* as green, despite the fact that where the C-conditions are fulfilled it *is judged* as red.

should use a slightly different criterion to pick out the same notion. His argument is that the basic equation fails to pick out what we were intuitively after; and so we should use the provisional equation instead. He is not after a different notion.

Because it's Wright's argument that I'm following, I'm going to stick with his equation; nothing of importance turns on this I think. The argument could be reformulated using basic equations; the choice between the two approaches will turn on how well they deal with examples like the shy and intuitive chameleon. But, since I think that the notion of an extension-determining concept just is the notion of a judgement-dependent echo concept, I'll treat his argument as an argument to the conclusion that intention is a judgement-dependent echo concept, and hence that it is, *a fortiori*, a response-dependent concept. This will enable me to apply the lessons we learned in discussing Pettit to what Wright has to say.[22]

IV WRIGHT ON INTENTIONS

Wright is concerned with "self ascriptions of psychological states like sensation, emotion, mood, belief, desire, and intention". In particular, he focuses on self ascriptions of intention. There are of course two rather different things that one could mean by the self-ascription of an intention: either the formation of a belief about one's intention, or the utterance of a sentence reporting that intention. Wright is primarily concerned with the former.[23] He is concerned with the possibility that the concept of intention is response-dependent: with the possibility that it is related to our judgements about our intentions in the same way that the concept of red is related to our judgements about what is red. This, Wright says, would be shown if the following provisional equation were *a priori* true:

$$C(\text{Jones}) \rightarrow (\text{Jones intends to } \phi \text{ iff Jones believes he intends to } \phi)$$[24]

There are two things to note about this. Firstly, the provisional equation is not concerned with the concept of intention quite generally, but with Jones' concept of his own intentions. Of course, if Wright's enquiry is to be of any interest we will need to know if the equation will still be *a priori* true if we substitute other people's names for that of Jones. But nevertheless, even if we show that the account applies to everyone, we will not have shown that the concept of intention in general is response-dependent. Rather, we will have shown that the concept that *each of us has of our own intentions* is response-dependent. Consider, in contrast, the concept that people have of other people's intentions:

---

[22]There are a number of other differences between the two types of concept, but I don't think that they are terribly important for our purposes. For a discussion see 'Objectivity Refigured', Appendix 3: On Two Distinctions.

[23]This is not immediately obvious, since Wright describes the relevant class of self-ascriptions as the class of avowals. But his discussion makes clear that his main focus is on the link between an agent's intentions and the agent's beliefs about his or her intentions.

[24]'Wittgenstein's Rule-following Considerations' p. 252. For the sake of consistency with the rest of my discussion I have reversed the order of the biconditional so that the response is on the right hand side; this does not, of course, affect anything of substance.

the second and third person concept as we might call it. To show that this concept is response-dependent we would need to show that the following equation is *a priori* true:

C (The Intention Assessors) → (Jones intends to ϕ iff The Intention Assessors believe he intends to ϕ)

where *The Intention Assessors* (the people whose judgements matter) include not just Jones, but all those who are in a position to form judgements on his intentions. Wright thus distinguishes first person ascriptions of intention, and the first person concept of intention that goes with it, from second and third person ascriptions, and the concept that goes with them. Separating off a first person concept of intention in this way might seem highly unattractive; and it is surely contrary to the Wittgensteinian spirit which Wright sees as his inspiration. Nevertheless, it is not an accidental feature of Wright's approach. What he is trying to do is to show how beliefs about our own intentions are special: how they differ from beliefs about others' intentions. He achieves this by proposing a distinctive first person concept of intention.

The second point to note is that what Wright has given us is a schema: an *intention schema*. This will give a basic equation for one of Jones' particular intentions when the variable ϕ is replaced with a verb phrase. In asking whether the first person concept of intention is response-dependent, Wright is asking whether each such instantiation will be *a priori* true. The thought appears to be that the concept of intention will be response-dependent if and only if the concept of each particular intention is response-dependent: the concept of intending to keep working for another hour, the concept of intending to take a bath, and so on.

But this is surely too strict a requirement. Remember that we saw that it is not satisfied in the case of colour concepts. We did not say that colours are response-dependent on the grounds that the schema

C(x) → (x is F iff x is judged to be F)

is *a priori* true if F is replaced with *any* colour concept, since it will not be true if F is replaced with the concept of extravagance. We said that the schema only yields an *a priori* true sentence if F is replaced with a colour concept which denotes a colour with which we are acquainted under that concept; only such colour concepts are response-dependent. Should a similar restriction be applied to the intention schema too? Is every particular intention concept response-dependent, or, for instance, only those with which we are acquainted?

There are two worries here, but the first is easily deflected. It concerns the opacity of ascriptions of intention, an opacity which does not seem to affect ascriptions of colour. Suppose I intend to abuse the driver of the car that has just cut in in front of me; but I do not intend to abuse my head of department. However, unbeknownst to me, the driver of the car that has just cut in in front of me is (of course) my head of department; so there is a sense in which I do intend to abuse my head of department. If this is right then perhaps there is a sense in which I intend to abuse my head of department without realizing that I do. But this is a familiar point, and poses no special threat here. We need simply to insist that we individuate intentions finely, so that the intention to abuse the driver in front of me is not the same as the intention to abuse my head of

department when I do not realize that the driver of the car in front of me is my head of department.

The second worry is more troublesome, and concerns the range of the variables in the intention schema. It parallels the difficulty raised by the concept of extravagance. I'll give an example. Several times recently someone has taken the windscreen wipers off my car when it is parked outside at night. I don't know who is doing this, nor why. It could be theft. It could be vandalism. But, since it only happens when the car is parked in one particular place, I am beginning to suspect that it is done by someone who objects to my parking there: perhaps because they don't like to have a car parked outside their house; perhaps because they want to park there themselves. Call the intention that leads them to remove my windscreen-wipers 'the tiresome intention'. As I've said, I don't know what intention this is; it might be the intention to steal them, or to have some fun, or to dissuade me from parking there. So it's not true that if I had the tiresome intention I would believe that I did, since I could have the tiresome intention without realizing that it was the tiresome intention. The concept of being the tiresome intention is clearly the concept of an intention; but it is not a response-dependent concept. So we need to restrict the intention schema so that it does not include intentions like the tiresome intention. How could we do this?

We saw in the case of colour that we could either just give a list of the response-dependent colour concepts, or else give a characterisation of them in terms of acquaintance. But the former is impossible in the case of intention: we cannot list them since there are indefinitely many intention concepts that we would expect to be response-dependent. So we are forced to the second characterisation. Again, however, there are problems. We surely cannot say that for an intention concept to be response-dependent we must be acquainted with that concept, since that makes the thesis much too limited. Every day we have intentions that we have never had before; any interesting account of intentions must apply to them. Equally we cannot say that it applies only to concepts which denote intentions which are such that if we had them we would know that we had them, since that would make the provisional equation trivially true. Perhaps there is a way to delimit the range of the variables in the intention schema, but I do not know what it is.

Having noted this problem, let's put troublesome concepts like that of the tiresome intention to one side and press on. What happens if we instantiate the intention schema with concepts which we clearly understand? Will this give us equations which are *a priori* true? To fix ideas, take one particular instantiation:

> C (Jones) → (Jones intends to resign from his job iff Jones believes he intends to resign from his job)

Suppose this equation is *a priori* true, so that Jones' concept of his intention to resign from his job is response-dependent. Then that concept will be a judgement-dependent concept: in believing that he intends to resign Jones judges it to be the case that he has that intention. Moreover, it will be an echo concept: the judgement on the right-hand side is a judgement about the applicability of the concept on the left. And it will be a users' concept: it is Jones who uses the concept of his own intention to resign, and it is Jones who is the judge of that concept. So Jones will be immune to error in his self-ascriptions of his intention to resign.

But that is surely not right. Jones could be in error about whether or not he has that intention. As Wright says, we are familiar with the idea that people can be wrong about their own intentions. We are familiar with the idea that they can be self-deceived. Jones might believe that he intends to resign from his job when in fact he intends no such thing. When it comes to the crunch Jones might discover that his desire for a regular income outweighs, and always has outweighed, his desire to be free of his loathsome job. He might discover that he never really intended to resign at all. The motivation for treating intention as a response-dependent concept was the hope that it would explain how it is that we are so accurate and so confident in our self-ascriptions of intention. We didn't want to rule out the possibility of error altogether. If the result of seeing intention as a response-dependent concept were to rule out the possibility of error, then that would present us with a reductio: we would be forced to conclude that it is not.

What should we do? Wright's response is to build a 'no-self-deception' clause into the C-conditions. The C-conditions will require that Jones has the concepts necessary to characterize his intentions, and that he is paying attention to them. To these Wright adds the further requirement that Jones is not self-deceived. The result is that we no longer have the unwelcome guarantee that Jones will be immune to error in his self-ascriptions of intention. That guarantee will only apply if Jones meets the further condition of not being self-deceived; and if he meets that condition the guarantee is not unwelcome. However, the proposed solution brings with it another problem: how do we formulate a no self-deception condition in such a way that it will not render the resulting provisional equation trivially *a priori* true? Remember our worry with triviality: if we allow whatever-it-takes formulations of the C-conditions we can arrive at an *a priori* true provisional equation for just about any concept. So our specific worry here is: how do we give a no self-deception condition which is not a whatever-it-takes condition? What can we say about self-deception other than that it is a state in which people do not know what they intend? As Wright states the problem:

> We seem to be very close to writing in a condition to the effect that the subject be 'free of any condition which might somehow impeded his ability reliably to certify his own intentions'.[25]

And that just amounts to the claim that the subject will believe he has a certain intention if and only if he does. Substituting the proposed C-conditions into the provisional equation gives us:

> Jones has the relevant concepts, is paying attention, and believes that he has a certain intention if and only if he does → (Jones intends to resign from his job iff Jones believes he intends to resign from his job)

which is trivially true.

It is important to see just what is at stake here. Part of the appeal of a response-dependent account of intention was that it seemed to offer an explanation of how it is that our beliefs about our own intentions are generally so accurate. But the first provisional equation which we constructed would have

---

[25]'Wittgenstein's Rule-following Considerations' p. 251.

had the consequence that we could *never* be wrong about our intentions. That cannot be right, since we are sometimes self-deceived. So Wright adds a further clause to the C-conditions, a clause which restricts the scope of the equation, and hence the immunity to error, to those subjects who are not deceived about their own intentions. But when he tries to formulate such a condition it leads to a trivial formulation of the provisional equation; and a trivial provisional equation provides no reason for thinking that the concept it contains is response-dependent.

How does Wright respond to this? Not by searching for a new formulation of the no self-deception condition, a formulation that will yield a provisional equation which is not trivially true. He thinks that no such formulation is available. Rather he embraces the trivializing formulation of the condition, the formulation which requires that the subject believes he has a certain intention if and only if he does. And he embraces the trivial provisional equation to which it gives rise. This appears to leave us with nothing, for we have said that no conclusion can be drawn from such an equation. Nevertheless, Wright thinks that there is a way forward. To see quite what this way is, we need to do some textual analysis. Here is what he says:

> We have, I think, to depart somewhat from the approach which emerged in the case of colour. But a possible variant of it is suggested by the reflection that the troublesome no-self-deception condition is *positive-presumptive*. By that I mean that, such is the 'grammar' of ascriptions of intention, one is entitled to assume that a subject is *not* materially self-deceived, or unmotivatedly similarly afflicted, unless one possesses determinate evidence to the contrary. Positive-presumptiveness ensures that, in all circumstances in which one has no countervailing evidence, one is *a priori* justified in holding that the no-self-deception condition is satisfied, its trivial specification notwithstanding. Suppose, then, that we succeed in constructing an *a priori* true provisional biconditional:
>
> C (Jones) $\rightarrow$ (Jones believes he intends to ϕ iff Jones intends to ϕ)
>
> where C includes the (trivial) no self-deception condition but no other trivially formulated conditions. Then if - lacking evidence to the contrary - we are *a priori* justified in holding the no-self-deception to be met, we are also *a priori* justified in believing the result of deleting that condition from the provisional biconditional in question. Likewise for any other positive-presumptive conditions listed under C. In this way we can eventually arrive at a restricted provisional biconditional in which all the C-conditions are substantially specified, and which, in the absence of any information bearing on whether the conditions are satisfied which we have deleted from it, it is *a priori* reasonable to believe.
>
> It is true that we are now dealing with something *a priori* credible rather than *a priori* true. But the question still arises: what is the explanation of the *a priori* credibility...[26]

---

[26]'Wittgenstein's Rule-following Considerations' pp. 251-2.

The explanation, Wright suggests, is that the concept of intention is response-dependent.[27] In giving his argument Wright introduces two technical terms. The first is that of *positive-presumption*; the second, that of *a priori justification*. I can see two ways of understanding what he means by them. I'll take them one at a time.

First construal

There are three kinds of evidential state that you can be in *vis-à-vis* a given proposition. You can have evidence for it; or evidence against it; or no evidence either way. A reasonable rule you might apply is the following: if the first state obtains, believe the proposition; if the second, disbelieve it; if the third, withhold judgement. On my first construal of Wright, a positive-presumptive proposition is one for which this simple rule doesn't apply: you're justified in believing a positive-presumptive proposition not only when you have evidence for it, but also when you have no evidence either way.

What of *a priori* justification? Given the way that I have just construed positive-presumption, I'm not sure what to make of it. An obvious gloss would be that we are *a priori* justified in believing P if and only if it is *a priori* true that we are justified in believing P. However, that cannot be what Wright means. For, in the passage just cited, he says that if a statement is positive-presumptive it will follow that when we have no evidence against it, we are *a priori* justified in believing it. If we gloss *a priori* justification in the way I have suggested, that is not right. Suppose we have no evidence against a certain positive-presumptive statement. We cannot conclude that in such a situation it is *a priori* true that we are justified in believing it. Given the way I have interpreted positive-presumption it might be an a posteriori matter whether or not the statement is positive-presumptive; moreover it will be an a posteriori matter whether we have evidence against the statement. So it will be an a posteriori matter whether we are justified in believing it.

How else can we understand the notion of *a priori* justification given this first construal of positive-presumption? I don't think we need to worry. The first construal of positive-presumption is clearly wrong, as we will see if we follow the argument through. Wright claims that the no self-deception condition is positive-presumptive. That is, he claims that the biconditional

X intends to F iff X believes that he or she intends to F

is positive-presumptive for any agent X. The central idea here is familiar enough: it's the idea of the burden of proof. Wright is suggesting that when it comes to self ascriptions of intention, the burden of proof is in the agent's favour. Is he correct in saying this? Surely he is. If someone believes that they intend to do a certain thing, that is good enough grounds for us to conclude that they do indeed so intend. We don't need further evidence before we come to that conclusion. The burden of proof is in their favour.

---

[27]Or, rather, that it is extension-determining. Remember that I am treating Wright's argument as though it were an argument to the conclusion that intention is response-dependent.

Let's accept that the no self-deception condition is positive-presumptive. That means that for each of Jones' possible intentions, we will be able to construct a (trivially) *a priori* true conditional of the form:

> Jones has the relevant concepts & is paying attention & has the belief that he has a certain intention if and only if he does → (Jones believes he intends to resign iff Jones intends to resign)

in which the third conjunct of the antecedent is positive-presumptive. Suppose we have no evidence for thinking that conjunct false; then we are justified in believing it is true. Now Wright argues that since the conditional is *a priori* true, we are in turn *a priori* justified (whatever that might mean) in believing that the result of detaching that conjunct will be true. He concludes that we can arrive at a non-trivial provisional equation which, under certain circumstances, we will be *a priori* justified in believing.

Putting aside the question of whether the argument is valid (this will depend on how we interpret *a priori* justification), we can ask whether it provides a reason for thinking that the concept of intention is response-dependent? The answer is no. There are many concepts for which we can construct similar arguments which are clearly not response-dependent. Consider the case of Jerry, who runs my local milk bar. He is good at identifying dollar coins. Give him a dollar coin saying it's a dollar coin and he checks and accepts it. Give him a dollar coin saying it's a two dollar coin and he shakes his head and gives it back. Give him a dollar coin for something that costs fifty cents and he gives you fifty cents change. We can construct a basic equation linking Jerry and dollar coins which is trivially *a priori* true:

> Jerry is paying attention & has the relevant concepts & is not deceived about the dollar coininess of x → (Jerry would believe that x is a dollar coin iff x is a dollar coin)

The equation is trivially true because the third conjunct of the antecedent is a 'whatever-it-takes' clause. But suppose I respond to this triviality as follows. The burden of proof is in Jerry's favour. If he says that something is a one dollar coin, I believe him, and I am quite justified in believing him, unless I have conclusive evidence to the contrary. The third conjunct is positive-presumptive. So I should be able to detach it, and I will be *a priori* justified (whatever that might mean) in believing the resulting, non-trivial conditional.

This doesn't provide evidence for thinking that Jerry's concept of a one dollar coin is response-dependent. I could have given a similar argument for just about any property which someone can reliably identify: squareness, for instance, or being made of gold. Clearly the first construal is wrong. The problem stems from the fact that the evidence I have for thinking that Jerry is reliable at identifying one dollar coins is a posteriori evidence. We need a way to construe Wright's words which does not allow this.

Second construal

According to the second construal, P is positive-presumptive if and only if *it is a priori true that* one is entitled to assume P unless one has determinate evidence to the contrary. Perhaps this is what Wright means when he says that it is a feature of the *grammar* of ascriptions of intention that they are positive-presumptive. Far fewer statements will be positive-presumptive in this sense than

were in the first sense. In particular, we will no longer be able to run the argument about Jerry since it is not *a priori true* that when it comes to dollar coin identification the burden of proof is in his favour.

Can we now make better sense of the idea of *a priori* justification? Somewhat. We might try to rehabilitate the idea that a statement is *a priori* justified if and only if it is *a priori* true that we are justified in believing it. If a statement is positive-presumptive then we are, in this sense, *a priori* justified in believing that if there is no evidence against it, it is true. To say this is just to restate that it is positive-presumptive. But the notion of *a priori* justification has wider application. Suppose we know *a priori* that A entails B. Suppose further that A is positive-presumptive. Then we know *a priori* that where we have no evidence against A we are justified in believing B. So the following statement is *a priori* justified: if we have no evidence against A, B is true.

Is this what Wright means by *a priori* justification? When he says a statement is *a priori* justified does he mean that it is *a priori* true that we are justified in believing it? I suspect he does; this is what I will take him to mean. But if so he makes an error in his reasoning. Remember that Wright says that if a statement is positive-presumptive, then it will follow that where we have no evidence against it we are *a priori* justified in believing it. If we understand *a priori* justification in the way that I have just suggested, this is simply wrong. The conclusion does not follow. In saying that it does Wright commits a scope error: the same error that is made when we infer from the fact that it is *a priori* true that if Jones is a bachelor he is a man, to the conclusion that if Jones is a bachelor then it is *a priori* true that he is a man. It is an a posteriori matter whether Jones is a bachelor, and it is an a posteriori matter whether he is a man. Similarly it is an a posteriori matter whether or not we have evidence against a positive-presumptive statement; and it is an a posteriori matter whether or not we are justified in believing it. What we can conclude from the fact that a statement is positive-presumptive is that where we have no evidence against it we are *justified* in believing it, not that we are *a priori justified* in so doing.

With our new definitions in hand let's turn again to the argument. It has three steps. First, Wright claims as a premise that the no self-deception condition is positive-presumptive. Second, he reasons from this to the conclusion that the provisional equation for intention, minus the no self-deception condition, is *a priori* justified. Third, he argues that this provides a reason for thinking that intention is response-dependent. I'll take the steps in turn.

First the premise. Is it really so clear that the no self-deception condition is positive-presumptive in this strong sense? Certainly some propositions are. Let me give an example. Suppose I pick someone from the population at random. I don't tell you who I've picked, so you don't know how tall they are. But it is *a priori* true that unless you have evidence to the contrary you will be justified in believing that they aren't in the shortest percentile of people. That statement will be positive-presumptive. (That was a little quick; in fact, whether you will be *a priori* justified in believing that depends on the epistemic rules that you feel yourself bound by. Certainly the odds are 100-1 in your favour, but you might not think that's good enough.)[28]

---

[28]Incidentally, the example is enough to show that one of the things Wright says cannot be true, namely that you can detach one positive-presumptive condition

Is the no self-deception condition positive-presumptive in this way? It's not clear. Isn't it possible that we have *learned* that people are generally reliable in their self-ascriptions of intention? Indeed isn't it conceivable that they might in fact not be? Do we know *a priori* that they are? I find it rather hard to get a grip on these questions. But I raise them as worries. Let's press on to see what Wright can conclude if we grant him the premise.

Suppose we have a conditional of the form ((A & B & C) → D). And suppose that C is positive-presumptive. Then it will follow that if we have no evidence against C, we are justified in believing the result of deleting C: that is, we are justified in believing ((A & B) → D). But we are not *a priori* justified in believing it. To think so is to commit the scope confusion that I spoke of earlier. Consider then the non-trivial equation that we get when we detach the no self-deception condition from the provisional equation, namely:

> Jones is paying attention & Jones has the relevant concept →
> (Jones believes he intends to resign iff Jones intends to resign)

Suppose that we have no evidence leading us to think that the no self-deception is not satisfied. Are we then *a priori* justified in believing this equation? No; we are simply justified in believing it. To think otherwise would be to commit the scope error. Wright claims that the non-trivial equation is *a priori* justified. But he is mistaken; he commits the scope error. The non-trivial equation is not *a priori* justified. Now we need to know what follows from his mistake.

Wright's case is based on the claim that although he cannot give a non-trivial *a priori* true provisional equation for intention, he can give an equation that is *a priori* justified. The idea is that that is good enough. To claim that we are *a priori* justified in believing the provisional equation is to give it a certain *a priori* status; and that suggests that there is some conceptual link between the intentions and the beliefs about the intentions. But we have seen that Wright does not give us grounds for thinking that the provisional equation is *a priori* justified. What he gives is reason for thinking that in certain circumstances we are justified in believing it. And that, surely, is not good enough. We have no reason to conclude from that fact that the notion of intention is response-dependent. Does this mean that Wright's case simply collapses? I think not. But we need to take a slightly different tack.

We need to examine some features of the no self-deception condition, features which the reader must have noticed, but which I have avoided commenting on till now. Recall how the condition is formulated: Jones believes he has a certain intention if and only if he does. Note first that this condition effectively renders the others redundant: circumstances where Jones is not paying attention or fails to possess the relevant concepts will be ones in which he fails to believe that he has an intention when he does. So the other conditions can be

---

after another in an a priori conditional, and still end up with something which you are a priori justified in believing. Indeed, you need not end up with something that you are even justified in believing. Suppose we have a conditional: If the person isn't in the first percentile, and isn't in the second, and isn't in the third... and isn't in the ninety-ninth, then they will be in the one hundredth. That is a priori true, and each of the conjuncts of the antecedent is positive presumptive. But if we detach each of the conjuncts we are clearly neither justified nor a priori justified in believing that which remains: namely, that the person is in the one hundredth percentile.

dispensed with. Second, note that as it is formulated the no self-deception condition is too strong. It requires that Jones not be wrong about any of his intentions. That is much too strict a requirement; it is doubtful that any of us would meet it. All that we require is that Jones is not wrong about the intention that we are currently analysing. Third, note that the no self-deception condition is just the provisional equation shorn of its C-conditions. Putting these points together we can see that the trivial provisional equation is really an equation of the form:

(Jones intends to resign iff Jones believes he intends to resign) →
(Jones intends to resign iff Jones believes he intends to resign)

What has Wright done to generate such an equation? He appears to have followed a procedure that we could use quite generally. Take any biconditional of the form

(F iff X believes F)

which is positive-presumptive. Construct from it a trivially true provisional equation of the form

(F iff X believes F) → (F iff X believes F)

(Further conditions can be added to the antecedent if you like; they make no odds.) Now delete the antecedent on the grounds that it is positive-presumptive. And conclude that it is *a priori* true that in circumstances in which you have no evidence against the antecedent you are justified in believing the consequent. Of course that is right if the antecedent is the consequent: we knew that all along The manoeuvre gets us nowhere new.

Nevertheless, seeing the manoeuvre in this way highlights something of interest. When Wright claimed that the no self-deception condition was positive-presumptive he effectively claimed that the provisional equation was positive-presumptive. And to claim that *is* to give the provisional equation a certain *a priori* status. We would do better to ask straight out: is positive-presumption evidence for response-dependence? In other words if we accept that it is positive-presumptive that Jones believes he intends to do something if and only if he does, is this in itself a reason for thinking that the concept of intention is response-dependent?

V RESPONSE-DEPENDENCE AND POSITIVE-PRESUMPTION

What is the relation between positive-presumption and response-dependence? I'll approach the question in a rather roundabout way. I'll argue by example. You can show the consistency of a theory by providing a model for that theory. I'll do something rather similar: I'll show how positive-presumption and response-dependence can be related by showing how they are related in one particular account of mind. The account of mind I'll take is that provided by analytic functionalism.[29] Unlikely as it may seem, I think that this theory does

[29]I have in mind a theory along the lines of that put forward by David Lewis in 'An Argument for the Identity Theory' and 'Mad Pain and Martian Pain' both in his *Philosophical Papers* Vol I (Oxford: Oxford University Press, 1983) Or rather I have in mind a theory which embraces the functionalist elements of the account

treat the first person concept of intention as response-dependent. I am not suggesting that, beneath his Wittgensteinian garb, Wright is really an analytic functionalist. But we might expect that the links that hold between positive-presumption and response-dependence in the context of analytic functionalism would also hold in the context of other theories of mind. And besides, analytic functionalism, or something like it, is so widely held that there is independent interest in seeing how the ideas we have been discussing apply to it.

Analytical functionalism provides us with a set of biconditionals which it claims to be *a priori* true. These biconditionals define what it is for someone to be in a certain mental state.[30] On the left-hand side of each biconditional a certain mental state is specified; and on the right, the functional role that is definitive of that state. The functional role describes how someone who possesses the mental state will behave in different circumstances; and it gives details of the further mental states that they will go into as a result of possessing the mental state that is being defined. Someone who possesses a certain mental state need not meet *all* of the conditions given by the biconditional for that state; but they must fit sufficiently many of them.[31] What kind of biconditional will be given for an intention? One characteristic feature of an intention is that a person who has it will believe that they have it: so that will be one of the conditions on the right-hand side. But it will not be the only such condition. Conjoined with it there will also be claims about the person's behaviour under various circumstances, and about their other mental states. So what we have is an *a priori* biconditional, linking, on one side, an intention with, on the other, a belief about that intention, behaviour under various circumstances, and other mental states.[32] Does this mean that the analytic functionalist will treat intention as a response-dependent concept? It does; but we will have to broaden our account of the ways in which concepts can be response-dependent.

The response-dependent concepts discussed so far have had basic equations in which the right-hand side has contained a single clause which has made reference to a single kind of response. However, not all response-dependent concepts are like this. Some have basic equations which contain a number of clauses on the right-hand side. Consider the concept of being sickly-sweet. Something is sickly-sweet if and only if it provokes *both* a feeling of nausea *and* the impression that it is sweet. Sickly-sweet has a basic equation that contains two clauses: the first concerns its sickliness, the second its sweetness. If a concept has a basic equation containing more than one clause, and if at least one

---

put forward there without embracing the identity theory. I think I could amend what I say here about functionalism so that it applied to a functionalist identity theory of the kind proposed by Lewis; but it would take me too far out of my way to do so.

[30]Or, rather, in Lewis's theory, they say what is *typically* the case for someone to be in a certain mental state. Since Lewis embraces the identity theory he can say that an agent has a certain belief if and only if he or she is in the physical state that *typically* realises the functional state associated with that belief.

[31]Again, on Lewis's account they need not meet any of them provided they possess the physical state which *typically* realizes the functional role for beings of their type.

[32]I've assumed that the functionalist will treat intentions as mental states; in fact they might prefer to analyse intentions in terms of beliefs and desires, and then give functionalist account of the beliefs and desires.

of those clauses makes reference to a response, I'll call it a *multiple clause* response-dependent concept.

This definition leaves open the possibility of multiple clause response-dependent concepts which contain some clauses that make no mention of responses. Here is an example: the concept of being a nauseous egg. Something is a nauseous egg if and only if it is both nauseous and an egg. Being a nauseous egg is a multiple clause response-dependent concept; its basic equation has two clauses, one of which makes reference to a response. I'll call such a concept an *impure* response-dependent concept: impure because although one of its clauses makes reference to a response, one does not. Some impure concepts are not so obviously impure. Take that of being an aphrodisiac. Aphrodisiacs arouse sexual desire. But not everything that arouses sexual desire is an aphrodisiac. A person isn't an aphrodisiac, however much desire they arouse. Aphrodisiacs are drugs; to be an aphrodisiac something has to arouse sexual desire *and* be a drug. So being an aphrodisiac is an impure concept.

The distinctions that we applied to single clause concepts, those between judgement concepts and non-judgement concepts, echo concepts and non-echo concepts, users' concepts and non-users' concepts, will not apply straightforwardly to multiple clause concepts. Rather than being features of multiple clause concepts simpliciter, these will be features of the *clauses* of the basic equations of those concepts. So, rather than asking whether a multiple clause concept is a judgement-dependent concept *simpliciter*, we must ask whether each of its clauses is a judgement-dependent clause; and the same goes for whether they are echo clauses or users' clauses.

What of the mental concepts proposed by analytical functionalism? These are multiple clause response-dependent concepts.[33] We can treat the biconditional associated with Jones' intention to resign as a basic equation for that concept. The right-hand side will have a clause making reference to Jones' belief that he has such an intention; and it will have other clauses making reference to other mental states of his and to his behaviour in certain circumstances. For instance, it will say that confronted with a propitious moment for resigning, Jones will resign. Will analytic functionalism entail that Jones is immune to error in his self-ascription of his intention to resign? That depends on the importance accorded to the various clauses. Suppose the analytic functionalist thinks that the only clause that really matters in the biconditional for Jones' intention to resign is the one that refers to his belief about that intention. Suppose she treats the others as so much window dressing: it is, she says, a necessary and sufficient condition of Jones intending to resign that he believes he intends to resign. If that were the way she saw things, then the crucial part of the biconditional would be simply the claim that Jones intends to resign if and only if he believes that intends to resign. So Jones' concept of his own intention to resign would be a judgement-dependent concept: it is a judgement that features on the right-hand side. And it would be an echo concept: the same concept feature on both sides of the equation. And it would be a users' concept: only Jones is the user of his concept of his intention to resign, and he is the judge of that concept. So if that were how the analytic

[33]Will they be impure? This will depend on how externalist the functionalist's account of mental properties is.

functionalist treated the concept of intention, Jones would be immune to error in his beliefs about his own intentions.

But that is not how the analytic functionalist typically sees things. Suppose Jones does believe that intends to resign. Suppose however, that he does not integrate this into his beliefs in any way. Suppose he makes no plans about what he will do after he has resigns, no plans as to how he will live when his income stops, no plans as to how he will spend his time. Suppose further that, faced with the opportunity to resign, he makes no move. Suppose he buys a new annual season ticket for the trip to work. Suppose, in other words, that Jones meets none of the clauses of the right-hand side of the functionalist's biconditional for the intention to resign except for one: that of believing that he intends to. In such a case the analytic functionalist will probably not ascribe to Jones the intention to resign. Jones does not meet the conditions of the biconditional closely enough to count as having that intention. So the analytic functionalist will probably not ascribe to Jones immunity from error about his own intentions. not say that Jones is infallible about his own intentions.

But what of positive-presumption? Might the analytic functionalist nevertheless think that it is *a priori* true that we are justified in believing the statement

Jones intends to resign iff Jones believes that he intends to resign

unless we have evidence to the contrary? Again it will depend on the importance of Jones' belief about his intention in the functionalist's biconditional for that intention. But suppose that on the functionalist's account Jones' belief is very important: so important that it will be true that Jones intends to do what he believes he intends to do unless *virtually all* the other criteria in the relevant biconditional indicate that he does not. In other words, suppose that Jones' belief about his intention is almost trumps, but not quite. If virtually all the other factors point the other way it will be over-ruled. Then it might well be that, where we have no other evidence, we will be justified in believing that Jones intends to resign just on the grounds that he believes he does. If we know that Jones believes he intends to resign, that is an extremely good reason for thinking that he does.

Moreover, the biconditionals proposed by the analytic functionalist were claimed to be *a priori* true. So it is reasonable to assume that the *relative importance* of the various clauses in those biconditionals will also be an *a priori* matter. If it is true that Jones' beliefs about his intentions are of central importance then it will be *a priori* true that they are. So if we are justified in believing that Jones intends to do what he believes he intends to except where we have evidence to the contrary, then it will be *a priori* true that we are so justified. In other words, the claim that Jones is accurate in his self-ascriptions of intention will be positive-presumptive.

I have had to make a few suppositions in order to paint the picture of analytic functionalism that I have painted: suppositions about exactly how the functionalist would treat intention. But they have not been outrageous suppositions. I have ended up describing a theory which ascribes to intention many of the features that Wright ascribed to it. It treats the concept of intention as response-dependent. And it treats statements like:

Jones intends to resign iff Jones believes that he intends to resign

as positive-presumptive. There is a further way in which the analytic functionalism I have described parallels Wright's approach. Wright treated the first person concept of intention in a different way to the second and third person concept. The analytic functionalist does the same. On the functionalist account there is a difference between *Jones'* concept of his intention to resign, and *our* concept of his intention to resign. The functionalist biconditional for Jones' intention to resign contains no clauses making reference to *our* beliefs about Jones' intention; only to *his* beliefs. So it provides no reason for thinking that a statement of the form

> Jones intends to resign iff people other than Jones believe that he intends to resign

will be positive-presumptive. To put the point another way, on the functionalist view I have been sketching, Jones' concept of his own intention is a users' concept: it is he who uses the concept, and it is his responses which figure in the relevant biconditional. It is in virtue of this fact (amongst others) that the biconditional which links Jones' intention with his belief about his intention is positive-presumptive. In contrast the second and third person concept of intention is not a users' concept: when I ascribe to you a certain intention I am ascribing to you something that is, on this view, quite independent of my responses. So there is no positive-presumptive biconditional linking your intention to my belief about your intention.[34]

So we have something of a vindication of Wright's claims. According to a popular theory of mind the concept of intention is response-dependent. You might by now be thinking that the best way of making sense of what Wright says is to treat him as an analytic functionalist. Not me; I would never think so rash a thing. All I think is that this is one way of accommodating his claims; I don't doubt that there are others. Perhaps there are some of a more Wittgensteinian bent. But I won't pursue that question. Instead I want to investigate what follows from our ability to make sense of the idea that intention is a response-dependent concept. I'll return to our starting point: to Wright's claim that a response-dependent account of intention provides us with an alternative to the Cartesian epistemology of mind with its notion of 'inner tracking'.

VI TRACKING THE MENTAL

According to Wright

> The Cartesian takes the authority of avowals [of intention] as a symptom of, as it were, a superlatively sure genre of detection.

---

[34]There could be an account which treated the second and third person concept of intention as a users' concept: which claimed, for instance, that it is a priori true that to have an intention is to be judged to have that intention, where the judge need not be the person who has the intention. Some have seen such an idea as providing the core of Davidson's interpretationism; others have seen it as the central idea in Dennett's notion of the Intentional Stance. I can't discuss such a theory here. Let me just say how it differs from Wright's approach. Wright's approach is one which seeks to use a response dependent account to explain what is *special* about first person knowledge: to explain how it *differs* from our knowledge about others' intentions. In contrast, this other approach claims that the very feature which Wright thinks is special about the first person concept is shared by the second and third person concepts.

> We should accomplish a very sharp perspective on the sense in which this is a 'grammatical' misunderstanding if it could be shown that avowals fail the order-of-determination test - in other words, that subjects' best opinions determine, rather than reflect what it is true to say about their intentional states, with the consequence that the notion of detection or 'inner tracking' as it were, is inappropriate.[35]

Put into our preferred idiom, Wright's point is this: if we can show that intention is a response-dependent concept, we will have shown that it is inappropriate to think of us *detecting* our intentions. Is this right?

Consider again the concept of red. Suppose we accept that red is a response-dependent concept. Does this entail that we are wrong if we treat our knowledge about what is red as the result of a kind of *detection*? Does it show that it is a mistake to think of us *tracking* the colour of things? Surely not. Our faculty of colour perception is a paradigm of the kind of thing for which such talk is appropriate; and this is something that the response-dependent account can accommodate. It can accommodate the idea that things really are coloured, and would have been even if we had not been around to see them; and it can accommodate the idea that we track the colours which objects have. The response-dependent account doesn't tell us how we do this; it's up to science to do that. And science obliges: it provides an explanation of how we track the colours of things, an explanation involving light and reflectance properties, and rods and cones and optic nerves.

Of course, the response-dependent account (at least as I have described it) does provide an *a priori* guarantee that we will not be wrong about which things are red. But what that shows is that we have arranged our concepts in such a way that the judgements we come to as a result of our tracking will not be false: it does not show that we do not track. In general the existence of an *a priori* connection between two concepts does not show that there cannot be a causal connection between the things to which those concept apply. To take an example of Armstrong's: there is an *a priori* connection between a thing being a poison and it causing illness or death in those foolish enough to consume it. But that does not preclude there being a causal link between the poison and the illness. Similarly, there is an *a priori* connection between an object being red and our judging it to be red. But that does not preclude there being a causal link between the object and our judgement, the kind of causal link which is described in the scientist's account of how we track things.

If the response-dependent account of colour is true, it does not show that we are mistaken in thinking that we track the colour of things. If it shows that we are mistaken at all it shows that we are mistaken at a higher level: mistaken in thinking that the *system of colour concepts* is itself something which we have simply read off the world. That is, it shows that our system of colour concepts is not dictated by nature - or at least, not by that bit of nature which excludes us. So it is here that according to the response-dependent account the notion of tracking is inappropriate: we might think that when we classify certain things together as red we are classifying them in accord with a scheme of classification which itself is something we have tracked. But it is not. The scheme is something which comes from our responses as much as from the world.

---

[35]Wright 'Wittgenstein's Rule-following Considerations' p. 250.

So much for colour; what of intention? The same point holds. There are two things that we might mean when we say that the notion of tracking is appropriately applied to our knowledge of our intentions. The first is that we track our intentions; the second is that we track the concept of intention itself. The response-dependent account of intention entails that the second of these claims is wrong. According to the response-dependent account the concept of an intention is not one that is dictated by the natural arrangement of our minds; that is, the concept is not dictated by an arrangement of our minds that holds independently of our beliefs about it. To that extent the idea of tracking the mental is out of place.

Perhaps the response-dependent account will force a revision to the Cartesian epistemology of mind insofar as the Cartesian picture includes the idea that the concept of intention is something we have tracked. However, the revision will not be very great, since the idea that we have tracked the concept of intention is not central to that picture. What is central to the picture is the idea that we track our intentions themselves. According to the Cartesian picture we track our intentions much as we track the colour of the objects around us. But as that comparison should make clear, this is something with which the response-dependent account has no quarrel. Just as the response-dependent account of colour is compatible with the idea that we track the colour of the objects around us, so the response-dependent account of intention is compatible with the idea that we track our intentions.

Of course, the response-dependent account of intention does not *demand* such an epistemology: it is quite compatible with an epistemology which makes no use of the idea of tracking. And it certainly says nothing about *how* we track our intentions: that is something which scientists would have to tell us. The response-dependent account is simply silent on whether the notion of tracking is appropriate or not.

What this shows is that Wright is looking in the wrong place if he wants something to undermine the Cartesian picture. The response-dependent account of intention, intriguing as it is, will not do the job. We can grant virtually all that Wright says about the response-dependent nature of first person ascriptions of intention, yet the essentials of the Cartesian picture remain untouched.