Intellectica, 2003, 36-37, pp. 227-246

William CHILD[↓]

Autonomy and Self-Knowledge

Abstract: Questions about autonomy play a key role in the debate between causal and non-causal views of action explanation. And these questions interact with questions about the nature and basis of self-knowledge. It is sometimes claimed (for example, by Wittgenstein) that the fact that our knowledge of our reasons is immediate and groundless shows that reasons cannot be causes of the beliefs and actions they explain. The paper argues against that claim. First, a model of self-knowledge is presented (itself derived from Wittgenstein) on which self-ascribing a belief involves converting a judgement that expresses that belief into a judgement that explicitly self-ascribes it. Then that model is adapted and applied to explain how we are able effortlessly and reliably to self-ascribe our reasons for our beliefs and actions - in a way that respects the immediacy of our knowledge of our own reasons whilst being entirely compatible with the view that reasons are causes.

Key words : autonomy, self-knowledge, self-ascription, reasons, causes, believes.

Résumé : Autonomie et connaissance de soi. Les questions qu'on peut se poser au sujet de l'autonomie relèvent de la controverse entre les conceptions causaliste et non causaliste de l'explication de l'action. Dans cette controverse, certains (p. ex. Wittgenstein) soutiennent que le fait que la connaissance que nous avons de nos raisons est immédiate et absolue prouve que ces raisons ne peuvent pas être les causes des croyances et des actions qu'elles nous servent à expliquer. En effet, la connaissance des causes est une connaissance empirique, laquelle ne saurait être acquise à la manière dont nous connaissons nos raisons. Ce chapitre s'oppose à cette façon de voir. Est d'abord avancé un modèle de la connaissance de soi (dérivé du même Wittgenstein) d'après lequel le fait de s'auto-attribuer une croyance implique la conversion d'un jugement exprimant cette croyance en un jugement qui l'auto-attribue explicitement. Ensuite, ce modèle est remodelé de façon à lui permettre d'expliquer comment il se fait que nous soyons capables, de manière fiable et sans effort particulier, de nous auto-attribuer les raisons de nos croyances et de nos actions. Ce modèle s'avère propre à rendre compte du caractère immédiat de la connaissance de nos raisons tout en restant

[↓] University College, Oxford OX1 4BH U.K.

parfaitement compatible avec l'opinion selon laquelle ces raisons sont des causes.

Mots-clés : autonomie, connaissance de soi, auto-attribution, raisons, causes, croyances.

1. INTRODUCTION

When philosophers consider our knowledge of our own minds, they are typically interested in our knowledge of *what* we believe, intend, expect, and so on. This knowledge is immediate; we have it without observation and without inference from our behaviour. And, though mistakes are possible, we are generally and reliably right about our own attitudes. The traditional philosophical question is, how is that effortless reliability to be understood? I want to raise a different question: how do we know *why* we believe that p, *why* we intend to M, or *why* we are acting as we are? How do we know our *reasons*? Our knowledge of our reasons shares the features of our knowledge of what we believe, intend and so on: it is immediate, non-observational and non-inferential; and, though we are fallible, we are reliably right about what our reasons are. But how do we know our reasons? What explains this aspect of our knowledge about ourselves?

This question about self-knowledge may seem distant from the topic of the autonomy of human action. But the two are closely related. Understanding human autonomy involves understanding the connection between our actions and our reasons for doing them; and a correct account of that connection must explain how we are able so effortlessly and reliably to know the reasons for our actions. Correspondingly, as we shall see, philosophers have drawn wide-ranging implications about the character of human autonomy from observations about the way in which agents know about their reasons for acting as they do.

2. AUTONOMY AND TWO POINTS OF VIEW

Philosophical discussions of human freedom have often exploited the idea that we can adopt two quite different points of view on human beings and their behaviour.[1] When someone performs an intentional action it is possible to describe what happens in terms that make no use at all of the concepts of agency, action, intention or reason. That is the description we get if we adopt an impersonal, or non-intentional, or objective point of view. From that point of view, we see the agent simply as a complex physical object; we describe what happens when she acts as the movement of parts of that object;

[1] See, for example, Strawson, 1966; Davidson, 1980; Dennett, 1971. The idea of seeing questions of freedom in terms of the relation between two points of view goes back at least to Kant's *Groundwork for the Metaphysics of Morals*.

and we explain those movements in terms of their causes in the agent's brain, nervous system and tissues. This description and explanation may or may not reveal that each effect is produced by a preceding cause in accordance with fully deterministic laws of nature. But whether or not determinism is true, the explanation achieved from this point of view says nothing about reasons or intentions; the concept of a free action, or of an agent's autonomy, does not figure at all. On the other hand, we can also adopt a personal, or intentional, or participant point of view. From this point of view we see the agent as a person - a creature with her own perspective on the world, with ends and purposes, who engages in reasoning about what to think and do. We see her behaviour as intentional action - the upshot of her own deliberation. And we explain those actions by giving the agent's reasons; we make her behaviour intelligible by showing how it was reasonable for her, given her beliefs and ends, to act as she did. The descriptions and explanations offered from this point of view presuppose some degree of human autonomy and, therefore, an ineliminable role for the agent herself.

Suppose we take it for granted that human behaviour is, under some description, completely explicable from the impersonal point of view, without reference to beliefs, intentions, and so on.[2] What that shows depends on our theory about how the accounts that can be given from the two points of view are related. We can distinguish three general kinds of theory.

A first kind of theory concludes that human freedom is an illusion. The true story about our behaviour is the one told from the impersonal point of view, by the physical sciences. The story about agents having beliefs and purposes and acting for reasons may for some purposes be a useful one. But ultimately it is only a *façon de parler*; human beings are not really the free agents we take them for.[3]

The second and third theories, by contrast, employ the framework of two different points of view to argue that human autonomy is compatible with the availability (in principle) of a complete, non-intentional explanation at the physical level. The second theory sees reason-explanation as itself being a form of causal explanation but argues that that fact does nothing to undermine our autonomy, even when seen in the light of the causal truths describable from the impersonal point of view. One version of this theory, for example,

---

[2] Though I make this assumption for the sake of illustrating the differences between different theories, it is not obviously correct; perhaps intentional notions play a part all the way down in physical explanation.

[3] This is the view taken by eliminativists like Churchland, 1981. It is also the view sometimes suggested by Dennett, who writes that descriptions achieved from the intentional stance are true only "with a grain of salt" or "only if we exempt them from a certain familiar standard of literality" (see Dennett, 1987), p. 72-3.

holds that the very same events can be explained by both intentional and non-intentional explanations, and that every event mentioned in an intentional explanation has its non-intentional aspect; but, it is argued, it follows from the unformalizability of the norms of rationality that there are no strict laws of action by reference to which human actions could be predicted or explained; and that, it is thought, shows that human agents are not mere creatures of deterministic physical law.[4]

The third theory denies that reason-explanations are causal explanations at all. Giving an agent's reasons for doing what she did allows us to make her behaviour intelligible by seeing it as intentional action. But this intelligibility does not involve showing why something happened by reference to its causal antecedents. Instead, it involves a kind of sympathetic understanding: seeing another's action in a light that allows one to see some point in what she did, or to fit it into a kind of pattern distinctive of human life.[5]

What is the impact of this kind of non-causal view of action-explanation on questions about human autonomy? On the non-causal view, something's status as a rational action has nothing to do with its causal antecedents. It is a matter, rather, of its position in a rational structure that involves the agent's other behaviour (actual and potential), the considerations she cites or would cite in its favour, and its context. As long as the appropriate structure is in place, there is a rational action. So no facts about the causal history of the bodily movements involved in acting (including the fact, if it is a fact, that those bodily movements are completely determined by their causal antecedents) could undermine the status of human behaviour as rational action. By the same token, there can be no question of vindicating the autonomy of human action - by showing, for example, that determinism is false or that there is some sort of neurophysiological correlate of autonomy. As long as human behaviour exhibits the appropriate rational structure, it is autonomous action. Of course, our behaviour can also be regarded from an impersonal point of view. But, on this way of seeing things, whatever the impersonal point of view reveals, it has no power to undermine the descriptions achieved from the personal point of view.[6]

---

[4] For this conception, see Davidson, "Mental Events", and "Freedom to Act".

[5] See, e.g., Anscombe, 1963.

[6] I have stated the non-causal view in its most radical form, according to which facts about the causal antecedents of human behaviour are simply irrelevant to its status as rational action. Less radical views would allow that some sorts of causal antecedent would disqualify behaviour from counting as rational action, whilst holding on to the claim that there are no particular positive requirements on the aetiology of rational action. For example, the non-causalist might say that certain kinds of robot that successfully mimic human behaviour do not really act intentionally; and he might trace

3. THE LINK BETWEEN AUTONOMY AND SELF-KNOWLEDGE

One issue in the debate about autonomy, then, is whether or not reason-explanation is a form of causal explanation. We can now appreciate the link between questions about our knowledge of our own reasons and questions about free action. For there is a line of thought that moves directly from the immediacy of our knowledge of our own reasons to a non-causal view of action explanation - with its attendant conception of autonomy. Crudely, we could present the argument like this:

> 1. Our knowledge of the reasons for our actions is immediate; it involves no observation or inference
>
> 2. It is impossible to have immediate knowledge of the causes of our actions
>
> 3. The reasons for our actions are not their causes

Such an argument is spelled out in Wittgenstein's *Blue Book*[7]:

> The proposition that your action has such-and-such a cause, is a hypothesis. The hypothesis is well-founded if one has had a number of experiences which, roughly speaking, agree in showing that your action is the regular sequel of certain conditions which we then call causes of the action. In order to know the reason which you had for making a certain statement, for acting in a particular way, etc., no number of agreeing experiences is necessary, and the statement of your reason is not a hypothesis.... - Giving a reason is like giving a calculation by which you have arrived at a certain result. (*Blue Book* 15)

Wittgenstein is surely right that my knowledge of my reasons for saying or doing something is not a hypothesis and is not based on experience of a past correlation between actions of this sort and preceding conditions of a given kind. But is he right to assume that knowledge of my reasons would have to be like that if reason explanations were causal explanations? What is his preferred account of our knowledge of our reasons? And what does he take that to imply about the character of reason-explanation?

---

that to the kinds of causal mechanisms that produce these robots' behaviour, without wanting to say that rational action is behaviour that is causally explained by the agent's reasons. (For examples of the sort of case that might provoke this reaction from non-causalists, see Block, 1981, and Peacocke, 1983 p. 205.) But the challenge for such views is to explain and motivate a constraint on how behaviour must *not* be caused if it is to count as rational action, without implying that rational action is, after all, behaviour that is causally explained by reasons.

[7] Wittgenstein, 1958a.

The main focus of this paper will be on the first of those questions: I shall argue that, though there may seem something paradoxical in the idea of non-observational, non-inferential awareness of a causal relation, a correct understanding of our knowledge of our own reasons is in fact perfectly compatible with a causal view of reason explanation. First, though, it is worth noting something else Wittgenstein himself said about our knowledge of our own reasons. In *Philosophical Investigations* §487,[8] he writes:

> "I am leaving the room because you tell me to."
> "I am leaving the room, but not because you tell me to."
> Does this proposition *describe* a connexion between my action and his order; or does it make the connexion?

Though Wittgenstein does not explicitly answer his question, commentators have assumed that the Wittgensteinian answer is that, "I am leaving the room because (or not because) you tell me to", *makes* the connection between my action and that reason rather than describing it.[9] That seems at least *prima facie* incompatible with a causal view of the relation between actions and the reasons for which they are performed. For suppose the difference between leaving the room for one reason and leaving it for another is a matter of what causally explains my leaving; then my statement of my reasons must surely be regarded as something that is made true (or false) by the causal relations that do in fact obtain - in other words, a description. Conversely, it seems that if my judgement about my reason *makes* the connection - makes it true that I am leaving for this reason rather than that - then the fact that I am leaving *because you tell me to* cannot be a causal matter; for how can my judgement about my reasons bring it about that my action has one rather than another set of causes? The immediacy and groundlessness of our identification of our own reasons may lead us to see our statements of our reasons as making rather than describing the connection between reason and action; but if we take that view, it seems that we must give up the idea that the reason-action connection is a causal one. That is, at any rate, the view that commentators have taken Wittgenstein's remarks to imply.[10]

4. SOME PHILOSOPHICAL ACCOUNTS OF SELF-KNOWLEDGE

---

[8] Wittgenstein, 1958b.

[9] See, for example, Hacker, 1996, for an exegesis of this and surrounding sections.

[10] In fact, I think that a causal view of reason-explanation can accommodate the idea that, in some cases, my judgement about my reasons *makes* the connection between a belief and my reason for believing it, or between an action and my reason for performing it. That will be so when, in making that judgement, I am reaching a (new) conclusion about what there is reason to believe or do. I discuss such cases very briefly in sections 6 and 7 below.

In considering our knowledge of our own reasons, it will help to get clear, first, about some simpler aspects of self-knowledge. How should we understand our knowledge of our beliefs, desires and intentions?

Suppose we accept the common-sense thought that, when we form a belief about our own attitudes, there is something there for us to be right or wrong about, independent of our forming that belief. Since we do by and large know what we believe, intend and so on, it seems that our self-ascriptive beliefs do succeed in tracking our attitudes. But how is that achieved? How do we reach our beliefs about our own attitudes; and what explains their reliability? We can review three philosophical accounts of self-knowledge, all of them, for different reasons, unsatisfactory.

First, there is the idea that all self-knowledge is based on introspected phenomenology - on feelings, sensations and experiences. That is the right model for knowledge of our sensations. But it does not generalize to belief, intention and the rest. Perhaps there are feelings that one often, or even typically, has when hoping that p, or when convinced that q, for example; but they are not essential to the attitudes themselves. Hoping, believing and the rest are simply not marked out by the way it feels to hope or to believe something.[11]

Second, there is the idea that ascribing attitudes to ourselves involves a kind of self-interpretation; we ascribe ourselves the beliefs and intentions that make best sense of our own behaviour. On this view, self-ascribing attitudes is an enterprise of fundamentally the same kind as the ascription of attitudes to others. But we are more reliable about our own attitudes, for two reasons: we have more experience of our own behaviour than other people's; and in our own case, the data for interpretation can include not just our actual behaviour and reactions but also our hypothetical reactions to possible situations contemplated in imagination. Now such a procedure of self-interpretation certainly plays a part in our knowledge of ourselves; that is how we recognize our possession of previously unconscious attitudes. But self-interpretation is clearly not the basic case; in most cases, there is no self-interpretation at all.

A third suggestion is that our self-knowledge is simply the result of a causal mechanism - a mechanism which, given an intention or belief as input, produces as output a belief in the subject to the effect that she has that intention or belief. So the immediacy of our beliefs about our own attitudes is the immediacy with which a causal mechanism produces its effects. And the reliability of those beliefs is

[11] Hume took just such a phenomenological approach to the attitudes. More recently, a view with some of these features has been advocated by Goldman, 2000.

the causal reliability of the mechanism that produces them.[12] Now there is something right in the idea that our attitudes reliably cause beliefs that we have those attitudes. When I believe that I M that p, that belief is susceptible of causal explanation. And the fact that I M that p presumably plays some part in the causal explanation. But even though it is *true* that our attitudes causally explain our beliefs about them, the appeal to a reliable causal mechanism does not by itself yield a satisfactory account of the epistemology of self-ascription. For it says nothing about what is actually involved, *from my own point of view*, in forming a belief about what I intend, expect or believe. Such beliefs are not based on introspected phenomenology or on an inference from my own behaviour. But that does not mean that I just find myself unaccountably possessed of beliefs about my attitudes - with no reason for those beliefs at all. Without an account of how we actually reach beliefs about our attitudes, of what self-ascription is like for the subject, the causal account is incomplete and, therefore, unsatisfactory.

So none of the three obvious ways of spelling out the epistemology of our self-ascriptive beliefs - observation, interpretation or a causal account - is satisfactory. That might be thought to imply that we must give up the common-sense idea that, in forming beliefs about our attitudes, we are forming beliefs about independently-existing states of affairs. But what is the alternative to that idea? Presumably, a version of the view that our introspective judgements are in some way constitutive of the attitudes whose presence they seem to report. As we have seen, that is a view that may be suggested by comments of Wittgenstein's in the *Blue Book* and *Philosophical Investigations*. And it has been developed by Crispin Wright.[13] It is an important view, which merits discussion. But I cannot examine it here. Instead, I want to describe a different view, also rooted in Wittgenstein, and endorsed by Gareth Evans. This view seeks to retain the common-sense thought that the beliefs, intentions etc. that we ascribe to ourselves are there anyway, independent of our self-ascriptions - while doing justice to the immediate, non-observational, non-inferential character of self-knowledge.[14]

---

[12] To avoid the implication that anyone who has a belief has an endlessly mushrooming series of higher- and higher-order beliefs, a plausible version of this view would need to modify the suggestion that beliefs *automatically* cause beliefs in their own existence. Perhaps my belief that p causes a belief that I believe that p only in response to the question, "What do I believe about p?"

[13] See, in particular, Wright, 1987; 1989; 199I; 1998.

[14] The view I describe in the next section is not original. Others who have developed views of self-knowledge with similar elements, often under the inspiration of Wittgenstein and/or Evans, include: Heal, 1994; Gordon, 1995; Hamilton, 2000; Moran, 2001. (Of course, these authors do not all agree with one another, or with me, in the way

5. A DIFFERENT ACCOUNT OF SELF-KNOWLEDGE

Consider first the case of belief. What one believes is a matter of how, from one's own point of view, the objective world is; believing that p is taking it to be true that p. Thus, when I make a judgement about how things are (the judgement that Oscar Wilde died in France, for instance) the explicit subject-matter of the judgement is the world outside me; but the fact that I make that judgement also gives information about me (the information that I believe that Oscar Wilde died in France). As Wittgenstein notes, we can exploit that fact; the point of getting someone to make a judgement may be precisely to get information about them:

> The language-game of reporting can be given such a turn that a report is not meant to inform the hearer about its subject matter but about the person making the report.
>
> It is so when, for instance, a teacher examines a pupil. (You can measure to test the ruler.) (PI pp. 190-1.)

When I report that Oscar Wilde died in France, I implicitly give information about myself. But suppose I want to give information about myself explicitly. To do that, I must move from a judgement that expresses what I believe to a judgement that explicitly ascribes that belief to me. And I do *that* simply by prefixing the judgement I am prepared to make about the external world with the clause, "I believe that". That turns a judgement about some aspect of the external world into a judgement about myself. But it does so without the need for any self-observation or for any investigation at all, over and above the assessment of evidence that has already gone into my judgement that things are, objectively, thus and so. Of course, it is not just my coming out with the *words* "I believe" that makes my report into an ascription of a belief to myself. I need to understand the words - i.e. to have the concept of belief; and that requires that I grasp the connections between belief and action, and that I think of belief as a property possessed not just by me but also by other people, each with their own perspective on a shared, objective world. But as long as I have the concept of belief, there is a simple recipe for ascribing beliefs to myself. To tell what I believe about where Oscar Wilde died, for example, this is what I do: first, consider the question, "Where did Oscar Wilde die?"; second, answer that question

---

they apply the basic Wittgensteinian thought.) What is new in the account I offer below is: first, the application of the basic Wittgenstein/Evans model of self-ascription to the self-ascription of *reasons*; and second (and relatedly), its use in resolving the problem posed by the accurate and effortless knowledge of what causally explains one's beliefs and actions.

- e.g. by judging, "Oscar Wilde died in France"; and, third, prefix that judgement with the clause, "I believe that".[15]

What does this account explain? I would stress four features. (i) The account answers the question, how from the point of view of the subject, beliefs about her own beliefs are reached. I do not just find myself possessing beliefs about what I believe; I *reach* those beliefs, by considering how things objectively are, understanding that the way things are from my point of view just is the way I believe them to be, and making the simple conceptual manoeuvre that turns a judgement that carries information about what I believe into a judgement that explicitly ascribes that belief to me. (ii) It explains the reliability of our self-ascriptive beliefs. The judgement I now make about p is already an expression or manifestation of my current belief about p; so a modification of that judgement is all that is needed to produce a correct self-ascription of the belief. Since I don't need to examine evidence, there is no room for error to slip in because of incompleteness in my evidence or mistakes in my assessment of it. (iii) Nonetheless, my belief that I believe that p does track an independent state of affairs; the higher-order belief is not constitutive of the first-order attitude. (iv) It is possible to be wrong about what one believes. How does the account I have sketched make sense of that fact? The topic deserves extended discussion. All I can offer here are a few pointers. The cases where one is wrong about what one believes will be cases in which one's explicit judgement about how things are does not express the belief that, deep down, one holds on the matter. So the belief one self-ascribes by operating the procedure I have described is the belief that corresponds to one's explicit judgement, rather than the belief that, deep down, one holds. Now in cases of successful self-deception, one really does have the belief one ascribes oneself: so the self-ascription of that belief is not mistaken; it is just that one also holds the opposite belief. But in some other cases, one does not at any level hold the belief one sincerely self-ascribes. In understanding these phenomena, we need a way of understanding self-deception, wishful thinking and other forms of irrationality. That is in itself an important philosophical task. But there is also the epistemological question, how our self-ascriptive beliefs can count as knowledge, given the possibility of mistakes of this sort. Even if the judgement I make about some aspect of things does express my real belief on the matter, how do I

---

[15] This Wittgensteinian position has been influentially endorsed by Gareth Evans: "in making a self-ascription of belief, one's eyes are, so to speak, or occasionally literally, directed outward - upon the world.... I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p.... If a judging subject applies this procedure, then necessarily he will gain knowledge of one of his own mental states" (Evans, 1982, p. 225).

know that it does? How do I know that I am not mistaken about what I believe?

There are various possible strategies for addressing this epistemological question. On one view, for example, the fact that my self-ascriptive beliefs are generally reliable is sufficient to ensure that, when true, those beliefs count as knowledge. On another view, what is necessary for knowledge of my beliefs is just that I reliably believe that I am making correct self-ascriptions when I am; it is not also required that I should reliably believe that I am subject to illusory appearances about my own beliefs when I am.[16] On a third view, knowledge of my own beliefs requires that I have some basis for the belief that my immediate self-ascriptions are by and large correct (perhaps from past experience that my immediate self-ascriptions are by and large congruent with my behaviour). There are clearly important issues here about irrationality and about epistemology. I cannot now set out a way of dealing with them. But I hope it is clear from the little I have said that the fallibility of the procedure I have described for self-ascribing beliefs need not be an insuperable barrier to the suggestion that that procedure gives us knowledge.

We have been discussing the self-ascription of belief. But the basic form of the account I have set out can be applied to the self-ascription of other types of attitude too. Consider the case of intention. When one intends to M, one commits oneself to M-ing. So there is a constitutive link between my answers to the questions, "What shall I do?", and, "What do I intend to do?". Correspondingly, the basic method for telling what one intends is to reach a judgement about what to do; one considers the pros and cons of M-ing, say, and judges, "M-ing is the thing to do", or, "I shall M". That is a judgement about what to do, or a judgement about M-ing; but the fact that I make it carries information about my intentions. I get from there to a judgement whose subject-matter is myself - a judgement that I intend to M - simply be prefixing my assessment of M-ing with the operator, "I intend to...". (This idea again has a clear Wittgensteinian pedigree.[17]) And we can give similar accounts for knowledge of what one expects, desires and so on.[18]

6. KNOWING THE REASONS FOR OUR BELIEFS

We have seen how we can make sense of self-ascriptions of current beliefs and intentions in a way that: (i) respects their immediate, non-observational, non-inferential character; (ii) accounts

---

[16] This mirrors a common response to Descartes' dreaming argument: to know that I am not dreaming, what is required is just that I should reliably believe that I am awake when I am; it is not also required that I should reliably believe that I am dreaming when I am.

[17] See e.g. *Philosophical Investigations* §588.

[18] For the Wittgensteinian roots of such an account for expectation, see e.g. *op. cit.* §586.

for their reliability; and (iii) does not merely appeal to a reliable causal mechanism, but gives a plausible account of what is involved in self-ascribing attitudes from the subject's own point of view. The account applies only to our knowledge of our current attitudes. Its essential element is that it is possible to be effortlessly right about one's current attitude by a conceptual manoeuvre that turns something that directly expresses that attitude into an ascription of the attitude to oneself. We cannot appeal to *that* source of reliability in explaining our knowledge of our *past* intentional states. For my present beliefs about my past attitudes to be reliable, there must be some kind of connection between my present inclination to judge, "I believed that p", and my past inclination to judge, "p". And that cannot be explained by seeing a judgement of the form, "I believed that p", as a simple modification of another judgement that directly expresses that past belief. For when I make the past-tense judgement, "I believed that p", the belief I am ascribing to myself is gone; there are no current judgements that immediately express that earlier belief.[19] So we have given only the beginnings of a comprehensive account of our knowledge of our attitudes.

However, we have enough to raise our next question: how do we know, not just *what* we believe or intend, but *why* we do so? How do we know our reasons? Suppose we take it for granted that reason-explanation is a form of causal explanation, and thus that knowing the reasons why someone believes what they do, or why someone is acting in the way they are, involves knowing what causally explains their belief or their action. How are we then to understand our immediate knowledge of our own reasons; how is it possible to have non-observational, non-inferential knowledge of a causal explanation?

I suggest that the account of self-knowledge that we have derived from Wittgenstein and Evans can take the causal aspect of reasons in its stride. Consider first the case of belief.

> (a) Jeanne believes that the Socialists will win the Presidential election.

Why does she believe that?

> (b) She infers that they will win the Presidential election from the fact that they won the mayoral election in Paris; that is to say, her reason for

[19] Of course, I may still believe what I believed in the past. In that case, there is one sense in which the belief expressed by my current judgement that p is the belief that I held in the past; *what I believe* now is the same as what I believed then. But there is another sense in which it is not: *my currently believing* that p is a different state of affairs from *my formerly believing* that p. And it is this latter sense of "same belief" that is the important one: my current judgement that p is a direct expression of my currently believing that p; but no current judgement of mine directly expresses my believing that p in the past.

> believing that the Socialists will win the Presidency is that they won Paris.

Reason explanation is causal explanation, so:

> (c) Jeanne's belief that the Socialists will win the Presidency is causally explained by her belief that they won Paris.[20]

That sets out the facts as they appear from a third-person point of view. Our question is, how does Jeanne come to know of that causal relation? And what explains her effortless reliability about it?

I think the story proceeds like this.

> 1. Jeanne believes that p.
>
> 2. Jeanne knows that she believes that p.

She gets this knowledge by operating the procedure for self-ascribing beliefs that we have already seen.

> 3. Jeanne asks herself: Why do I believe that p? What is my reason for believing that p?

This is a question about herself and what causally explains her believing that p.

> 4. Jeanne considers the question: What reason is there to believe that p?

Here, Jeanne is setting out to answer a question about herself; but she does so by thinking, in the first instance, not about herself but about the reasons for believing that p - about what makes it right or reasonable in the circumstances to believe that p.

> 5. Jeanne judges: the reason for believing that p is that q (i.e. the fact that q is, in the circumstances, sufficient reason to believe that p)

The explicit subject matter of that judgement is the reasons there are for believing that p. But, as before, the fact that Jeanne makes that judgement tells us something about her - that she believes that the fact that q is sufficient reason to believe that p.

> 6. Jeanne judges: my reason for believing that p is that q (i.e. I believe that the fact that q is, in

---

[20] Is Jeanne's belief that the Socialists will win the Presidency caused by *her belief that* they won Paris? Or is it caused by *the fact that* they won Paris; or by *her knowledge that* they won Paris? In this case, since her belief that the Socialists won Paris is a case of knowledge, and is caused by the fact that they did, all three causal claims are true. But, in general, a belief that q may cause a belief that p without its being the case that the fact that q caused the subject's belief that p; that will be so when the subject's belief that q is false. For present purposes I do not think anything turns on the question whether, when S's belief that q is true and knowledgeable, we should say that S's belief that p is causally explained by the fact that q or by her knowledge that q, rather than by her belief that q.

> the circumstances, sufficient reason to believe that p).

Here Jeanne moves from an answer to the question, "What are *the* reasons for believing that p?", to an answer to the question, "What are *my* reasons for believing that p". (Equivalently: she moves from an answer to the question, "Why *should I* believe that p?", to an answer to the question, "Why *do I* believe that p?".) That involves moving from the judgement, "The reason for believing that p is that q", to the judgement, "I believe that the reason for believing that p is that q". And Jeanne can make that move in the same way as she makes any other move from judging, "p", to judging, "I believe that p".

> 7. The judgement Jeanne makes at 6. is a case of knowledge.

This is just as in the standard case of self-ascribing a belief.

> 8. To believe "the reason for believing that p is that q" is (or involves) a disposition, in the circumstances, to infer p from q.

Belief in general involves dispositions to behave in ways that are appropriate given that belief and one's other attitudes. So, in particular, beliefs about what is, in the circumstances, sufficient reason to believe something involve dispositions to form beliefs on the basis of other beliefs. So:

> 9. Jeanne's effortless, authoritative knowledge that she believes that the reason for believing that p is that q *is* knowledge of what causally explains and sustains her belief that p.

That is the basic account of how we know the reasons why we believe what we do - an account that is completely compatible with the view that knowing the reasons for our beliefs is knowing what causally explains them. There is much more to say. For now, I will simply register three features or corollaries of the account.

(i) As Wittgenstein notes, there may be a difference between my original reasons for coming to believe that p and what I now take to be the reason (or reasons) for believing it:

> The question: "On what grounds do you believe this?" might mean: "From what are you now deducing it (have you just deduced it)?" But it might also mean: "What grounds can you produce for this assumption on thinking it over?" (*Philosophical Investigations* §479)

The procedure I have described is a procedure for telling what you now take to be good reasons for believing that p, and therefore (in my view, if not in Wittgenstein's) for telling what currently causally sustains that belief. Knowledge of one's original reasons for coming

to believe something (if any[21]) will involve memory; and an account of our authority about our own past reasons will depend on the account we give more generally of our past attitudes. That is for another occasion. But it is clear that thinking about the reasons for believing that p may lead one to see new reasons for something one already believed, or to abandon beliefs that one comes to see are unfounded. So the facts about what causally sustains one's current beliefs can be changed by deliberation. (In such cases, we could say that my judgement about my reasons makes the connection between my belief and my reason for it - and even that my judgement both makes and describes that connection.[22])

(ii) Sometimes we believe things for bad reasons. In that case, the considerations I give as my reasons for believing that p, considerations that cause me to believe that p, should not do so. The account I have given accommodates that easily. What underpins the fact that my belief that p is causally explained by my belief that q is not that q *is* a good reason for believing that p, but that I *believe* it is. If q is a bad reason for believing that p, then I am wrong to believe that p on the basis of q; but it remains the case that that is why I believe it - that my belief that p is causally explained by my belief that q.

(iii) Can we be wrong about the reasons why we believe something? If so, can the account I have given explain that possibility? I think we can be wrong about our reasons, but that the possibility of mistakes about *why* we believe something is no more problematic than the possibility of mistakes about *what* we believe; indeed, some mistakes about *why* I believe that p just are mistakes about *what* I believe is a good reason to believe that p. (One such case would be this. Pierre's wife is having an affair. He believes that she is faithful, and judges that the facts that p and that q are sufficient reason for believing that. But this judgement does not express his real belief about the goodness of his reasons; he really believes that the facts that p and that q are very flimsy reasons for thinking that his wife is faithful. So, when he moves from a judgement about reasons to the self-ascriptive judgement, "I believe that the facts that p and that q are sufficient reason to believe that my wife is faithful", his self-ascription is false. Consequently, he is wrong about what causally explains his belief in his wife's faithfulness.)

There are at least two questions to be asked about such mistakes. How do they arise; how should we understand the sorts of

---

[21] "If any" registers the fact that it is possible to acquire beliefs simply by picking them up - without acquiring them on the basis of reasons. It is plausible that, in learning our first language, that is how we acquire a great mass of framework beliefs.

[22] For the suggestion that a judgement may both make and report an intentional connection, see *op. cit.* §§ 682-4 (though Wittgenstein is there discussing past-tense judgements).

irrationality they involve? And what effect does the possibility of such mistakes have on our ability to know why we believe what we do? These are serious questions, which need to be answered. For now, I simply record my belief that, as with the parallel questions about mistakes concerning our own attitudes, it will be possible to answer them without giving up the basic model I have offered of our knowledge of our own reasons for belief.

7. KNOWING THE REASONS FOR OUR ACTIONS

I turn, finally, to our knowledge of the reasons for our own actions. If we take it for granted that my reasons for M-ing causally explain my M-ing, then how should we understand our non-observational, non-inferential, authoritative knowledge of causal relations? The basic model will stress, as before, that the way I know why I am doing something is not by looking for introspective evidence of a causal relation or by observing my behaviour and attempting to find the best interpretation of it, but by considering what reason there is for me to M. If I conclude that, in the circumstances, the reason for M-ing is that p, then I can move immediately from there to the claim that p is my reason for M-ing - the reason why I am M-ing. This is entirely consistent with the idea that my reasons for M-ing causally explain my M-ing.

Consider Wittgenstein's case.

(a) I leave the room.

Why do I leave? What is my reason for leaving?

(b) I am leaving, not because you tell me to, but to catch a train that goes at 18.00.

But reason explanation is a form of causal explanation. So, if 2. is true, then:

(c) My leaving the room is causally explained by my believing that the train goes at 18.00.[23]

But how do I know what *is* my reason for leaving the room and, therefore, what causally explains my doing so? We can set things out step by step.

1. I know that I am leaving the room

For present purposes, we may take this for granted.

2. I ask myself: Why am I leaving? What is my reason for doing so?

That is a question about myself and about what causally explains my leaving.

---

[23] As in the case of reasons for belief, there is an issue about whether my leaving the room is causally explained by *the fact that* the train leaves at 18.00, by *my knowledge that* it does, or by *my belief that* it does. That is an important issue. But, as before, we can bracket it for present purposes.

> 3. I consider the question: What reason is there to leave the room? Why should I leave?

As before, I set out to answer a question about my reasons not by thinking about myself and my attitudes but, in the first instance, by thinking about what reason there is to leave the room.

> 4. I judge: the reason for leaving the room is that the train goes at 18.00 - i.e. the fact that the train goes at 18.00 is the reason to leave the room (and the fact that you told me to leave is no reason).

At this stage, I make a judgement about *the reason* for leaving - about *why I should* leave. To know *why I am* leaving - what *my reason* is for leaving - I need to get from a judgement that expresses my belief about the reason for leaving the room to a judgement that explicitly ascribes that belief to me. So:

> 5. I judge: my reason for leaving is that the train goes at 18.00 - i.e. I believe that the fact that the train goes at 18.00 is the reason to leave.

My judgement at step 4 automatically *expresses* my belief about the reason for leaving. I get from there to a judgement, at step 5, that is *explicitly about* my belief about the reason for leaving by the usual Wittgenstein/Evans procedure for reaching beliefs about my own beliefs. And, since this judgement about my reason for leaving is not based on introspective or behavioural evidence, it cannot go wrong through my having insufficient evidence or making mistakes in assessing it.

> 6. Judgement 5. is a case of knowledge.

The procedure for getting from a judgement that p to a judgement that I believe that p is not just reliably truth-preserving; it yields knowledge.

> 7. Believing, "the reason for leaving the room is that p", is (or involves) a disposition to leave the room if, and because, one believes that p.

So:

> 8. My effortless, authoritative knowledge that my reason for leaving is that the train goes at 18.00 is knowledge of what causally explains my leaving.

We saw in the case of belief that deliberation may change my reasons for believing something, or lead me to form new beliefs and abandon old ones. Something similar is true for practical reasoning. Considering the question, "What reason have I for doing M?", may make me see new reasons for doing something I already intend to do (or am already doing), and it may lead me to abandon an intention I now see to be unsupported by my reasons. In these circumstances we can say, compatibly with a causal view of reason-explanation, that a

judgement about my reasons for acting makes the connection between an action and the reason for which it is done (whilst also saying that the judgement describes that connection).

Just as our knowledge of our reasons for believing what we do is not infallible, so our knowledge of the reasons why we do what we do is not infallible. In one sort of case, I am wrong about what my reasons for M-ing are: e.g. I think that I am making a donation because the recipients need help; in fact I am making it so that people will think well of me. In such a case, the problem is that my judgement about what is, in the circumstances, the reason for making a donation does not reflect my real, underlying belief about the reasons. So this is a special case of the phenomenon of fallibility about what one believes. As before, the investigation and explanation of this and other sources of error about our reasons for acting as we do is important. But I do not think it affects the tenability of the basic model I have offered.

8. CONCLUSIONS

The problem with which we began was how to understand our effortless authority about our reasons for believing what we do and acting as we do. As we saw, Wittgenstein was tempted to move from observations about the character of our awareness of our reasons to a non-causal view of the relation between reasons and actions. That had a knock-on effect on his conception of human autonomy.

My discussion has started from the case of knowledge of our own attitudes in general, and moved on to the case of knowledge of our reasons. What I have argued is: (i) there is no reason why the phenomenon of effortless, authoritative knowledge of our own attitudes should make us abandon the ordinary, realist view that when we ascribe beliefs, intentions, etc. to ourselves we are forming beliefs about attitudes that are there anyway, independent of our beliefs about them; the Wittgenstein/Evans model of the self-ascription of belief, which is extendable to intention and other attitudes, shows how the character of our self-knowledge can be accommodated within such a common-sense view; (ii) our knowledge of *why* we believe and act as we do is explicable on the same basic model as our knowledge of *what* we believe and intend; so, finally (iii) the fact that we can know the reasons for our actions immediately, without observation or inference, does nothing to support the non-causal strand in philosophical treatments of human autonomy.[24]

---

[24] This paper is a revised version of my presentation at a colloquium on the autonomy of the human agent organized at the Collège de France in March 2001 by Jean-Luc Petit, under the direction of Alain Berthoz. I am very grateful for the invitation to participate, and for the stimulation provided by discussion at the colloquium.

References

Anscombe, G. E. M. (1963). *Intention,* Blackwell, Oxford.
Block, N. (1981). Psychologism and Behaviorism, *Philosophical Review.*
Churchland, P. (1981). Eliminative Materialism and the Propositional Attitudes, *Journal of Philosophy.*
Davidson, D. (1980). Mental Events and Freedom to Act, *Essays on Actions and Events,* Oxford University Press, Oxford.
Dennett, D. (1971). Intentional Systems, *Journal of Philosophy.*
Dennett, D. (1987). Instrumentalism Reconsidered, *The Intentional Stance,* MIT Press, Cambridge, Mass.
Evans, G. (1982). *The Varieties of Reference,* Oxford University Press, Oxford.
Goldman, A. (2002). Simulation Theory and Mental Concepts, *Simulation and Knowledge of Action,* eds. Jérome Dokic and Jöelle Proust, John Benjamins Publishing Company, Amsterdam.
Gordon, R. M. (1995). Simulation Without Introspection or Inference from Me to You, in *Mental Simulation: Evaluations and Applications*, eds. Martin Davies and Tony Stone, Blackwell, Oxford.
Hacker, P. M. S. (1996). *Wittgenstein, Mind and Will: An Analytical Commentary on Wittgenstein's Philosophical Investigations,* vol. iv, Blackwell, Oxford.
Hamilton, A. (2000). The Authority of Avowals and the Concept of Belief, *European Journal of Philosophy*.
Heal, J. (1994). Moore's Paradox: A Wittgensteinian Approach, *Mind*.
Moran, R (2001). *Authority and Estrangement: An Essay on Self-Knowledge,* Princeton University Press, Princeton.
Peacocke, C. (1983). *Sense and Content,* Oxford University Press, Oxford.
Strawson, P. F. (1966). "Freedom and Resentment", *Freedom and Resentment and Other Essays,* Methuen, Londres.
Wittgenstein, L. (1958a). *The Blue and Brown Books,* Blackwell, Oxford.
Wittgenstein, L. (1958b). *Philosophical Investigations,* 2nd edition, eds. G. E. M. Anscombe, R. Rhees and G. H. von Wright, tr. G. E. M. Anscombe, Basil Blackwell, Oxford.
Wright, C. (1987). On Making Up One's Mind: Wittgenstein on Intention, *in* P. Weingartner and G. Schurz eds., *Logic, Philosophy of Science and Epistemology,* Hölder-Pilcher-Tempsky, Vienna.
Wright, C. (1989). Wittgenstein's Rule-Following Considerations and the Central Project of Theoretical Linguistics, *in* A. George ed., *Reflections on Chomsky,* Blackwell,Oxford.
Wright, C. (1991). Wittgenstein's Later Philosophy of Mind: Sensation, Privacy and Intention, in K. Puhl ed., *Meaning Scepticism,* de Gruyter, Berlin.

Wright, C. (1998). Self-Knowledge: the Wittgensteinian Legacy, in C. Wright, B. Smith and C. Macdonald eds., *Knowing Our Own Minds,* Oxford University Press, Oxford.