

PERSONAL AND SUBPERSONAL: A DIFFERENCE WITHOUT A DISTINCTION

The idea that there is a hierarchy of levels of psychological explanation is well-established in philosophy and cognitive science. Daniel Dennett, for example, distinguishes between explanation from the intentional stance at the top of the hierarchy, beneath which is explanation from the design stance and then explanation from the physical stance (Dennett 1987). Zenon Pylyshyn has a comparable hierarchy of semantic, functional and biological levels of explanation (1984). Other theorists have called the highest level of explanation the 'knowledge' level, or the 'ecological' level. I prefer to talk of the 'personal' level of explanation, and instead of a 3-way distinction I will be discussing a more straightforward distinction between the personal level of explanation and the subpersonal (although there are, of course, different levels and types of subpersonal explanation).

The burden of this paper is that, although personal states are importantly different from subpersonal states, these differences do not warrant the demarcation of an autonomous domain of explanation. To the extent that the hierarchical picture is committed to construing personal-level psychological explanation as autonomous it needs to be treated with caution. Although appeal to personal-level states is indispensable in rendering intelligible the behaviour of persons, there is no autonomous and self-contained domain of personal-level explanation.¹ Alternatively put, we must resist the inference from the fact that there are useful and indeed indispensable explanations of behaviour that appeal to personal-level states to the conclusion that personal-level states have their home only in an autonomous domain of explanation completely insulated from the subpersonal level.

§1 The model

It is widely held that there is a distinctive mode of explanation appropriate only to the behaviour of persons. Explaining the behaviour of persons is often held to be a distinctive type of explanation – for three reasons:

¹ I am neutral on the question whether there is a genuine distinction to be drawn between interpretative explanation in the 'Verstehen' mode and causal-nomological explanation in the 'Erklären' mode, as is frequently maintained by philosophers working in the hermeneutic tradition. But if the argument of this paper is sound it follows that explanation in the Verstehen mode, if such there be, cannot be identified with a level of explanation defined exclusively over subpersonal states.

- 1) Distinctive vocabulary It involves appeal to intentional states – to beliefs, desires and so on.
- 2) Distinctive laws It picks out classes of behavioural regularities that cannot be picked out at other levels of explanation.
- 3) Distinctive constraints It operates according to various principles of rationality – ascribing intentional states on the assumption that the person to which they are ascribed is rational.

The first of the three distinguishing features of personal-level explanations can initially be appreciated through a couple of examples.

Some philosophers place considerable stress on the cognitive significance of possessing a cognitive map, where the notion of a cognitive map is defined as "a representation in which the spatial relations of several distinct things are simultaneously represented" (Evans 1982, p.151). Possession of a cognitive map in this sense is often thought to be a vital element in a subject's understanding of the objectivity of the spatial environment, and indeed of his being self-conscious. This nexus of ideas ultimately goes back to Kant's Critique of Pure Reason. In this first sense of cognitive map, possession of a cognitive map is a high-level cognitive ability, something whose attainment in childhood marks a significant ontogenetic step. It is a form of knowledge: knowledge of spatial relations and knowledge of different routes between places.

But there is another important sense in which the notion of a cognitive map is deployed. In this second sense, cognitive maps refer to the storage of geometric information in the nervous system. Here is a recent definition from Gallistel:

A cognitive map is a record in the central nervous system of macroscopic geometric relations among surfaces in the environment used to plan movements through the environment. (Gallistel 1990, p.103)

Like the first sense of 'cognitive map', we are dealing here with the simultaneous representation of spatial relations. But the suspicion that these spatial relations are not being represented in the same way is confirmed when we read on in Gallistel's The Organization of Learning and find that all animals from insects upwards possess similar types of cognitive maps in this second sense – the cognitive maps that control movement in animals all preserve a system of metric relations within earth-centred coordinates. This is clearly something very different from the first sense of 'cognitive

map'. And it is a difference that one might capture by saying that 'cognitive map' is a personal-level term when used in the first sense, and a subpersonal-level term when used in the second sense.

As a second example, consider the state of looking at a particular object – say a horse – and recognising what sort of an object it is. The concepts that I possess lead me to classify an experience in a certain way. The result is a perceptual belief that I see a horse. Second, consider David Marr's theory of visual information processing (Marr 1982). According to Marr, visual information processing involves associating shape descriptions derived from the visual image with stored shape descriptions and 3-D models. This is also a form of visual classification, but apparently not of the same type as the other – classification of the second type can occur without classification of the first type (just as one can have a cognitive map of the second kind without a cognitive map of the first kind). This is precisely the sort of difference that might be captured by saying that the first state (the conscious recognitional state) is a personal-level state, while the second state (the state of the visual processing system) is a subpersonal-level state.

The second and third distinguishing features of personal-level explanation are both true of what is often called common sense psychology, or folk psychology. Common-sense psychology is what we use to manage our social interactions – to predict what other people will do in given situations, to work out what their reasons are for acting in the ways that they do act, and to speculate about what they will think that we think their reasons for acting are. Common-sense psychology is the psychological analogue of naive physics, or intuitive physics. Our grasp of commonsense psychology helps us to navigate through the social world, just as our grasp of naive physics help us to navigate through the physical world.

Nonetheless, personal-level explanation cannot be equated with commonsense psychology. Personal-level explanation includes much more than commonsense psychology. In fact much of cognitive psychology proceeds at the personal level. Reaction times, learning curves, patterns of error and verbal reports are all personal-level phenomena – as are the psychological principles that might be put forward to explain them. And when young children are given spatial tasks to test whether they do in fact represent space in the way that might be described as involving possession of a cognitive map, this is empirical work at the personal level. We are dealing with the familiar

common-sense psychological concepts of knowledge, belief and desire, but they are being experimentally refined and sharpened.

§2 The autonomy of personal-level explanation

It is widely held that explanations of behaviour at the personal level form a distinct and autonomous level of explanation. This is a view common both to many philosophers and to a dominant paradigm within cognitive science, although they argue for it in different ways and put it to different uses. I have already drawn attention to one very relevant idea here, which is that personal-level explanation makes use of and identifies true generalizations about human behaviour which cannot be captured at any other level of explanation. We can only make sense of various patterns that are revealed in human behaviour if we explain them in terms of personal-level cognitive states. Explanation goes hand in hand with counterfactual truths about what would have happened had circumstances been different, and there are true counterfactual claims which can only be captured by the generalizations of personal-level explanation. These generalizations are true in virtue of rational connections between content-bearing representational states. The general picture here is very familiar.

What is distinctive in claims of the autonomy of personal-level explanation is contained in the following two theses

- a). explanations at the personal-level can be fully understood without knowing any facts at the subpersonal level.
- b). subpersonal states will not feature in explanations of behaviour at the personal level.

In philosophical terms, these are most closely associated with Ludwig Wittgenstein and Gilbert Ryle. The thought is that when one tries to explain personal-level states in terms of states which are not themselves 'personal-level' states one is making some sort of 'category mistake'. Explanations of events at the personal level proceed in terms of other events at the personal level, and moving outside the personal level is changing the subject.

Those who believe in the autonomy of personal-level explanation stress the idea that personal-level explanation runs out at certain points. And in fact this is entailed by the first strand in the autonomy thesis. Explanation has to run out if the only psychological states that can be appealed to

are personal-level states. There is no personal-level state that explains why we feel an itch, for example. And this is something that emerges very clearly in the following passage from Dennett:

When we have said that a person has a sensation of pain, locates it, and is prompted to act in a certain way, we have said all there is to say within the scope of this [personal-level] vocabulary. We can demand further explanation of how a person happens to withdraw his hand from the hot stove. . . [but] if we do this we must abandon the explanatory level of people and their sensations and activities and turn to the sub-personal level of brains and events in the nervous system. But when we abandon the personal level in a very real sense we abandon the subject matter of pains as well. . . for our alternative analysis cannot be an analysis of pain at all, but rather of something else – the motion of human bodies or the organization of the nervous system. (Dennett 1969, 93-4)

I think Dennett's position becomes somewhat clearer if we make a distinction between two different types of explanation. In this passage Dennett is discussing horizontal explanation, roughly definable as the explanation of a particular event or state in terms of antecedent events or states. The autonomy theory is a thesis about horizontal explanation. It says that when human behaviour is being horizontally explained at the personal level, the only psychological states appealed to will be personal-level states. But when autonomy is understood in this sense it remains perfectly compatible with what I want to call vertical explanation.

Horizontal and vertical explanations have different aims. The horizontal explanation of personal-level psychology is strategic and predictive – it helps us navigate the social world. The vertical explanation of cognitive science, on the other hand, is not really strategic or predictive. It is legitimatory, in a sense which comes out very clearly in the following passage from Fodor:

One can say in a phrase what it is that computational psychology has been proving so successful at: viz. the vindication of generalisations about propositional attitudes, specifically, of the more or less commonsense sorts of generalizations about propositional attitudes. Thus, for example, we have got fragments of a theory of perception, and it makes clear how a computational system could regularly come to believe that P in causal consequence of its being visibly the case that P. . . what a computational theory does is to make clear the mechanism of intentional causation; to show how it is (nomologically) possible that purely computational – indeed, purely physical – systems should act out of their beliefs and desires. (Fodor 1986, p.422)

The vertical explanations of cognitive science are not alternatives to the horizontal explanations of personal-level psychology, and nor of course are they a part of it. Rather, they try to explain how the generalizations of personal-level psychology can be true.

At the risk of simplification, it is useful to distinguish two general strategies for legitimacy vertical explanations. One strategy aims to find, for any personal-level state, a subpersonal state which realises that state. This is familiar from the various forms of classical functionalism. The general project is to find a topic-neutral specification of the functional role of a given personal-level state and then to identify the subpersonal-level state which realises that functional role. A second strategy seeks vertical explanations, not of the individual personal-level states, but human and psychological capacities which are responsible for the individual personal-level states. This is the project of homuncular functionalism. The aim is to analyse those capacities and mechanisms into ever-simpler capacities and mechanisms until one eventually reaches operations that can plausibly be taken to be instantiated at the neural level.

§3 **Neuropsychology and the personal-subpersonal distinction**

Whichever broad strategy is adopted for providing vertical links between personal-level states and subpersonal-level states, the general picture espoused by believers in autonomy is that horizontal explanations at the subpersonal level explain the enabling conditions of personal-level explanation. Subpersonal explanation shows how personal-level explanation is possible, but cannot replace or fully capture what goes on at the personal level. This conception of subpersonal explanations as purely enabling provides an obvious passport to autonomy. One enabling condition of intentional action, for example, is that the laws of physics should hold, but the fact that the laws of physics hold will not feature in an explanation of any particular intentional action. And, one might argue, the same holds for the subpersonal background conditions – since they can legitimately be taken for granted they have no place in the horizontal explanation of particular intentional actions.

Difficulties with this view emerge when we start to consider the cognitive disorders resulting from damage to the brain that are studied by neuropsychologists. The various forms of amnesia are familiar examples, as are the reading deficits found in dyslexic patients and the speech difficulties found in aphasic patients. Other neurological deficits are more recondite, like blindsight (where patients who have no visual experience in an area of their visual field are nonetheless capable of discriminating stimuli of which they report no awareness) and unilateral neglect (where patients

'neglect' the left side of objects, including their own bodies, as they perform everyday tasks like eating and getting dressed).

Neuropsychology seems to present us with examples of personal-level states which cannot be given autonomous personal-level explanations, because what is going on at the subpersonal level ceases to be just a matter of background conditions. Blindsighted patients, for example, make visually-based responses, like reaching towards a presented stimulus, in the absence of any visual awareness of that stimuli (Weiskrantz 1986). How can this be given any sort of personal-level explanation? It is a vital part of the ordinary personal-level explanation of reaching actions that they involve an awareness of what one is trying to reaching. Such explanations will be variations on the theme that the person saw that something which they wanted was at that place. But this sort of explanation is obviously not available in the case of blindsighted patients. The appropriate explanation will clearly bring in subpersonal states, as the following passage makes clear:

Our everyday intuition is that a reaching action requires an awareness of what one is trying to touch or grasp. The study of blindsight shows that this is not necessarily so. . . In the case of blindsight, it seems most plausible that information transmitted from the input is failing to arrive at some higher-level subsystem, but the subsystems to which it can arrive are sufficient to effect appropriate reaching behaviour. It is simple to assume that it is the failure of the input to arrive at these higher level sub-systems that is responsible for the patient's lack of visual awareness of the stimuli. (Shallice 1988, p.388)

This seems clearly to breach the explanatory autonomy of the personal level. The personal-level action of reaching is explained in terms of visually derived information reaching subpersonal sub-systems that are capable of controlling movement, despite the fact that no information reaches those sub-systems which subserve visual awareness.

There seems to be an obvious reply here, however. Subpersonal facts are best viewed as necessary but not sufficient conditions of normal behaviour. This would be one way of explaining why they should be excluded from ordinary personal-level explanation. And it is perfectly compatible with the thought that they can provide sufficient conditions for a breakdown in performance – and so have a role to play in explaining abnormal, disordered behaviour.² So, according to this reply, there's nothing mysterious about the fact that neuropsychological disorders need subpersonal explanations. A defender of autonomous personal-level explanation will perhaps

² A position like this is considered in Taylor 1964 Ch. 1, particularly n.1 on pp.24-5.

withdraw the global claim that no subpersonal state can feature in the horizontal explanation of a personal-level state, while insisting that subpersonal states cannot feature when personal-level states are explained in the manner distinctive of personal-level explanation. The distinctive mode of explanation referred to here is, of course, explanation in which considerations of rationality play a constitutive role.

All those who believe that explanation of personal-level states is autonomous in the sense I have been discussing believe that the crucial factor marking off the personal level as a distinct level of explanation is the role of rationality. The point emerges clearly in the following passage from Pylyshyn:

In a cognitive theory, the reason we need to postulate representational content for functional states is to explain the existence of certain distinctions, constraints and regularities in the behaviour of at least human cognitive systems, which, in turn, appear to be expressible only in terms of the semantic content of the functional states of these systems. Chief among the constraints is some principle of rationality. (1984, p. 38)

A similar point is made by Donald Davidson:

If we are intelligibly to attribute attitudes and beliefs, or usefully to describe motions as behaviour, then we are committed to finding, in the pattern of behaviour, belief and desire, a large degree of rationality and consistency. (1980, p.237)

So, a defender of the autonomy of personal-level explanation will place a lot of weight on the distinction between normal and abnormal behaviour, and considerations of rationality will mark the difference. The blindsight point will be met by restricting the autonomy of personal-level explanation to explanation that can be assessed according to what one might broadly term criteria of rationality.

§4 Rationality: Explanation and prediction

The proposal to take rationality-involving personal-level explanations as primary places a heavy burden on the notion of rationality. The nature, function and role of these considerations of rationality clearly needs some scrutiny.

A good way to start is by setting out what seem to be some basic and undeniable facts about the social practices of personal-level explanation and prediction (taking prediction to be the complement of explanation).

(a) The standard form of a personal-level explanation is 'A ϕ -ed because $P_1 \dots P_n$ ' where $P_1 \dots P_n$ report certain personal-level psychological states of the subject. The implication is clearly that $P_1 \dots P_n$ rationalise or make comprehensible why A should have ϕ -ed. But the exact way in which they do this is almost never made explicit.

(b) We understand such explanations when they are made by others. We rarely have to ask how $P_1 \dots P_n$ explain A's ϕ -ing

(c) The standard form of a personal-level prediction is 'A will ϕ because $P_1 \dots P_n$ ' where $P_1 \dots P_n$ report certain personal-level psychological states of the subject. As with explanations, the implication is clearly that $P_1 \dots P_n$ rationalise or make comprehensible why A should have ϕ -ed. But the exact way in which they do this is almost never made explicit.

(d) We adjust our behaviour to the behaviour of others in a way that suggests that we have tacit abilities to predict their behaviour.

(e) We are capable of coordinated activity which depends on our all making the same predictions without consulting each other

The clear implication of (b), (d) and (e) is that we have what is often termed a shared "theory of mind".³ The clear implication of (a) and (c) is that this shared "theory of mind" issues in personal-level explanations and predictions that have a common form, differing primarily in tense. We need to look a little at this common form. In particular, what is the unexpressed but nonetheless predictive/explanatory relation between A's ϕ -ing (past or future) and A's being in states $P_1 \dots P_n$?

One canonical way of looking at this relation, going back at least to Aristotle, is put very clearly in the following passage from Donald Davidson:

If someone acts with an intention then he must have attitudes and beliefs, from which had he been aware of them and had he had the time, he could have reasoned that his act was desirable. . . If we can characterise the reasoning that would serve, we will, in effect, have described the logical relations between descriptions of beliefs and desires, and the description of the action, when the former gives the reasons with which the latter was performed. We are to imagine, then, that the agent's beliefs and desires provide him with the premises of an argument. (1980 pp. 85-86)

According to this classical view of prediction/explanation, the tacitly known theory of mind mentioned above can be understood in part as involving tacit knowledge of a range of inferential principles such that the suitable application of those principles to $P_1 \dots P_n$ will yield a description of

³ I am using the term in a way that is intended to be neutral between 'theory-theory' and simulationist accounts. I'm also steering clear of any strong claims about just how wide the 'we' ranges here.

A's \emptyset -ing. It is because we all have tacit knowledge of such principles that we don't need to make them explicit – that we can understand each other's explanations and predictions and act in coordination without explicitly going through and comparing the moves that would take A from P_i . . . P_n to the act of \emptyset -ing.

Underlying this general picture is a perceived symmetry. The possible transformation of P_i . . . P_n according to the relevant set of inferential principles into a description of A's \emptyset -ing is explanatory/predictive because it recreates A's own decision-making processes, which themselves proceed in accordance with the same set of inferential principles. Again, the details of the actual transformations involved in A's decision-making processes are no more frequently made explicit than are the details of the reconstructive transformations involved in the explanation or prediction of A's \emptyset -ing. The tacit nature of the inferential moves that take both A from P_i . . . P_n and the explainer/predictor from A's being in P_i . . . P_n to A's \emptyset -ing cannot be over-stressed.

But the tacit nature of these inferential moves brings with it an explanatory burden. We will not be able to understand personal-level explanations/predictions unless we understand the implicitly known inferential principles which govern them. This is a point that has not been sufficiently appreciated (at least by philosophers). It is frequently assumed, I think, but rarely made explicit that these inferential principles include the basic canons of what Ramsey called the logic of consistency. Again, this is an idea that goes back to Aristotle (with some variation in the appropriate canons of inference). The inferential principles implicit in explanation, prediction and practical decision-making are, on this view, such familiar deductive principles as modus ponens, modus tollens, contraposition and so on, in conjunction with such basic principles of probability theory as that strength of belief should be proportional to the degree of evidence for a particular proposition; that the probability of a conjunction can never be greater than the probability of its conjuncts; that the probability of a hypothesis and the probability of its negation should add up to 1; and so on. The state to which these inferential principles are applied are assumed to be governed by familiar requirements of deductive closure.⁴

⁴ For a stimulating discussion of exactly how to understand the requirements of deductive closure and consistency see essay 3 in Levi 1997.

To return to the first strand in the thesis of the autonomy of personal-level explanation, it is easy to see how it would follow that explanations at the personal-level can be fully understood without knowing any facts at the subpersonal level if the inferential principles presupposed by and tacitly applied in psychological prediction and explanation, as well as practical decision-making, were the familiar principles of the logic of consistency. If psychological predictions and explanations really are just condensed arguments which take personal-level states as premises and transform them according to these familiar inferential principles then there would indeed be no room for the subpersonal to get a foothold – all that we can expect from the subpersonal level is an account of how basic principles like modus ponens and modus tollens are hard-wired into the brain. This, one might think is all that is needed. We don't need an explanation or a justification of modus ponens. How, after all, could we provide such a thing without using modus ponens? All we need is a legitimating account of how modus ponens is implemented in the brain.

The problem with this picture is quite simply that the normative principles of rationality, as found in the propositional and predicate calculi and the probability calculus, are not adequate to the task of explanation and prediction. A fortiori they cannot be the inferential principles presupposed by and tacitly applied in psychological explanation and prediction. There are both a priori and a posteriori reasons for denying what might be termed the explanatory adequacy thesis. The a posteriori reasons all stress the large-scale divergences between the demands of normative theories of rationality and the actual reasoning abilities of human beings operating with limited information (both about the world and about themselves) and within a limited time-frame. Since normative theories of rationality are descriptive only of the reasoning habits of ideally rational beings, the explanatory adequacy thesis fails because it appeals to normative standards that do not properly describe the behaviour that is being explained or predicted. The a posteriori reasons should be sufficiently obvious and familiar to set the problem up. I'll return to the a priori reasons below.

There are, broadly speaking, three different strategies for dealing with the apparent difficulties for prediction/explanation posed by the mismatch between the demands of rationality and the reasoning abilities of human beings. One strategy is to hold that the demands of rationality are best viewed as idealizations, although a different type of idealization from, say, the gas laws. The

standards of rationality represent ideals to which rational agents aspire and, even when we can see agents conspicuously failing to live up to the standards of rationality, we can nonetheless make sense of their behaviour by viewing them as striving to satisfy those ideals. A second standard way of dealing with computational shortcomings is to revise the normative theory of rationality to produce a descriptively adequate model of "bounded rationality" (Simon 1982). A third strategy is to make a sharp distinction between viewing the norms of rationality as appropriate for self-criticism and the control of first-person deliberation and viewing them as tools for prediction and explanation. It is this third strategy I favour, but first some comments about the other two.

To appreciate how the first strategy might work, let's start from the platitude that we cannot live up to demands of consistency imposed by the normative theory of rationality. It is a demand of rationality, for example, that consistency be maintained among one's beliefs and the deductive consequences of those beliefs. Clearly, though, we know only a tiny fraction of the deductive consequences of our beliefs, and even within that tiny fraction we are computationally incapable of establishing consistency. Mechanical methods of establishing consistency, like truth-tables, rapidly become unusable as the number of atomic propositions is increased, since if n is the number of atomic propositions, then the number of lines to be checked is 2^n . Often quoted in this context is Cherniak's calculation that a computer capable of checking one line of a truth-table in the time it takes a light ray to cross a proton would need 20 billion light years to check a truth-table for a belief set that contained 138 atomic beliefs (Cherniak 1986).

Of course, the computational shortcomings of human beings will come as no surprise, and nobody has suggested that the predictive/explanatory adequacy of assumptions of rationality depends upon treating agents as ideally rational. Philosophers like Davidson who believe in some version of the principle of charity need not be too perturbed by this. On the plausible assumption that there will be relatively little divergence in different individuals' capacities to detect inconsistency and establish consistency, the workings of something like the principle of charity can be easily accommodated. When I consider whether or not I am attributing to a third party a set of inconsistent beliefs I can do no more than assess them for consistency by my best lights. The third party cannot reasonably be expected to do more than assess them for consistency by his best lights. Given that

assessments for consistency by my best lights are likely to yield much the same results as assessment for consistency by his best lights, my attribution to that third party of a set of propositional attitudes is unlikely to breach the requirement that it attribute a set of attitudes which would be inconsistent by the subject's own lights. Here we might indeed say that I am making sense of the third party by interpreting him as striving after the normative ideal of deductive consistency.

Still, some independent understanding is required of what the mechanisms for detecting inconsistency and establishing consistency actually are. It seems clear that truth-tables will not be much help. Twenty billion years is twenty billion years, and the notions of aspiration and ideals do not seem at first sight to get much of a grip here. Some account needs to be given of how these mechanisms (whatever they are) stand with regard to the normative ideal of maintaining consistency over the deductive closure of one's belief set. Further complexities emerge when one reflects that failures to maintain consistency are not all due to computational factors – or to other factors affecting 'performance' like distraction, lack of memory and so on. Self-deception, repression and weakness of will can be powerful forces. No theory that does not take them into account will be adequate for explanation/prediction. And yet how can one incorporate the prevalence of self-deception, repression and weakness of will into a predictive/explanatory theory while still representing agents as striving to abide by the norm of consistency?

The first strategy comes under more threat, though, from the well-documented experimental evidence that basic and apparently incontrovertible principles of deductive and inductive logic are regularly breached, even by sophisticated thinkers:

- A study carried out in 1977 (Rips 1983) showed that the only basic conditional argument that their subjects could apply reliably was modus ponens. There was a noticeable tendency to affirm the consequent and deny the antecedent. 21% of the subjects said that an argument which denied the antecedent would always be valid – while the figure was 23% with affirming the consequent. Also quite remarkable was the fact that 43% failed to see that modus tollens arguments were always valid.
- Another good example of failure in elementary deductive reasoning is to be found in the selection task experiments carried out by Wason and Johnson-Laird (1972). The subjects were presented with four cards (as pictured) and then asked to evaluate the conditional 'if there's a circle on the left then there's a circle on the right' by saying which cards they would have to see completely in order to answer the question.

INSERT WASON FIGURE

The answer, of course, is that cards (a) and (d) must be unmasked. Unfortunately, only 5 out of 128 college students realised this. Almost all the 123 who got it wrong failed to see the need to turn (d) over. This is failing to see the equivalence of a conditional, $p \rightarrow q$, with its contrapositive, $\sim q \rightarrow \sim p$.

- A significant factor in probability judgments is the prior probability of the relevant outcomes (base-rate frequency). Suppose I am given a character sketch of a particular individual within a given group and asked whether it is more likely that that individual is an engineer or a farmer, then my judgment ought, if it is minimally rational, to reflect what I know about the relative numbers of engineers and farmers in that population. Kahneman and Tversky (1973) tested this by giving subjects personality descriptions of individuals from a population of 100 lawyers and farmers and asked whether those descriptions were more likely to be of lawyers than of farmers.⁵ The subjects were divided into two groups, one of which were told that the population contained 30 lawyers and 70 farmers, while the other was told that it contained 70 lawyers and 30 farmers. Both groups came up with more or less the same answers.
- Another basic principle of probability judgments is that the probability of a given sample being representative of the population from which it is drawn varies in proportion to the size of the sample. Again, this basic principle is regularly contravened. One way in which it is contravened is by subjects tending to neglect that a large sample is less likely than a small sample to stray from the average of the population as a whole. Another way occurs because people have a tendency to apply the so-called 'law of small numbers' – ie to judge even small samples to be highly representative of the population as a whole (Tversky and Kahneman, 1971).
- A fundamental principle of the probability calculus is that the probability of a conjunction cannot be greater than the probability of one of its conjuncts.⁶ But subjects regularly commit the conjunction fallacy of assigning a higher probability to a conjunction than to one of its conjuncts (Tversky and Kahneman, 1982).
- It might also be worth mentioning the well-known Monte Carlo fallacy, where gamblers assume that a long run of similar outcomes at even odds will make an opposite outcome more likely.

It is clear that these deviations from what one might term the classical conception of rationality are not random. They arise because the reasoning strategies that human beings employ do not map straightforwardly onto the ideal reasoning strategies of the logician. Tversky and Kahneman argue, for example, that some of the deviations from the normative requirements upon probabilistic reasoning just noted arise because people use what they term a representativeness heuristic which is not sensitive to many of the factors that affect probability. The selection task has also been a rich source of speculation about the inferential principles that people actually employ in conditional reasoning (Evans and Over 1996 Ch. 4). On one interpretation, subjects are applying pragmatic reasoning schemas, rather than misapplying rules of deductive inference (Cheng and Holyoak

⁵ A sample description was: 'Dick is a 30 year old man. he is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well-liked by his colleagues'. This is obviously neutral between lawyers and farmers. Moreover, all the subjects correctly calculated the odds of an arbitrary undescribed individual being a lawyer, given the proportions in the population.

⁶ This follows, of course, from the so-called extension rule of probability – that if the extension of A includes the extension of B then $P(A) \geq P(B)$ – since the extension of the conjunction A & B is included in the extension of B and in the extension of A.

1985). On another interpretation subjects are seeking to reduce their uncertainty (calculated in information-theoretic terms) about whether or not the conditional is true (Oaksford and Chater 1994).

Recall that the first strategy for dealing with the mismatch between the normative demands of rationality and the actual reasoning abilities of human beings is to view the demands of rationality as normative ideals which agents can plausibly be represented as striving to attain, even though they cannot in any sense be expected to achieve them. So, we can try to make sense of the behaviour of others on the assumption that they are following, albeit imperfectly, the normative demands of the theory of rationality. But the experimental evidence shows that this is based on an erroneous conception of human reasoning. The strategies that we follow in real-life, real-time reasoning situations are not imperfect applications of the techniques and inferences prescribed by the normative theory of rationality, simply because they are not applications of those techniques and inferences at all. There is nothing incoherent in the idea that we might strive to realise unrealisable ideals (with religious vows a case in point). But one does at least have to be striving in the right direction.

Proponents of so-called "bounded rationality" have taken this lesson to heart and proposed that we revise our normative standards of rationality to accommodate it (Simon 1982). According to Simon and others, the principle of "ought" implies "can" requires us to effect a radical down-shift in the demands that we place on rational thinking and rational decision-making. Our normative standards must be tailored to the actual reasoning habits of normal human subjects. This approach seems misconceived, however. The "ought" implies "can" principle certainly rules out unrealisable ideals which we cannot even strive to realise. But if we can strive to realise those ideals then the "can" clause seems satisfied. It is no objection that we often do not strive to realise those ideals. The point is simply that we can (and sometimes do) strive to realise them. Descriptive inadequacy is not a barrier to a theory's being normatively binding, provided that subjects are capable of bringing the norms of the theory to bear on their own decision-making. But the descriptive inadequacy of a normative theory is a barrier to its serving a useful predictive/explanatory function.

This brings us to the third strategy for dealing with the mismatch between the normative demands of rationality and the reasoning abilities of ordinary agents. With very few exceptions philosophers have assumed that the normative theory of rationality performs both prescriptive and explanatory/predictive tasks – and that each stands or falls with the other. Thus, the theorist of bounded rationality concludes from the predictive/explanatory inadequacy of the normative theory of rationality that such a theory cannot serve a prescriptive role. The third strategy rejects this assumption and holds instead that the normative theory of rationality should be viewed as providing prescriptive principles which rational agents can reflectively employ to control and police their own deliberations, even though it is generally not the case that those very same agents employ the same prescriptive principles in their non-reflective decision-making (Levi 1997). A fortiori, the normative principles of rationality cannot generally be employed for the purposes of explanation/prediction.

That something like this version of the third strategy must be true follows from the particular ways in which the first and second strategies have been found wanting. It is further strengthened by an a priori argument based on considerations put forward by both Schick (1979) and Levi (1997). Both Schick and Levi identify what might be termed the paradox of the rational self-knower (my phrase not their's). The paradox is (roughly) this. A rational agent confronting a particular set of options cannot coherently predict that he will rationally choose a particular one of those options. Insofar as he believes himself to be rational, is aware of his preference-ranking, believes that he will adopt a particular choice-policy and believes that his choice will be effective he cannot consistently believe that he will choose a particular option. He'll be able to identify the option determined by his preference ranking and choice-policy (assuming that there is a single such option) and, believing himself to be rational, he must hold that he cannot but carry that option out. But if he cannot but carry that option out there is no sense in which he can be described as choosing it. The point is this. Prescriptive principles of rationality are employed in choosing a uniquely admissible option (or set of options) from a set of feasible options. But for agents who believe themselves to be rational the set of admissible options coincides with the set of feasible options. So the normative standards of rationality cannot in any meaningful sense be applied by such agents (to themselves).

The conclusion to draw is that we must drop what Levi calls the “smugness assumption” that the deliberating agent will choose rationally – if the norms of rationality are to be meaningfully self-applicable.⁷ But the “smugness assumption” has to be held if the norms of rationality are to be used for the purposes of explanation and prediction. We cannot assume that the normative principles of rationality are explanatory/predictive unless we assume that agents will make rational choices. Nor, of course, can we hold ourselves entitled to assume that agents will make rational choices in every case except when we are the agents in question.

Combining the a priori and a posteriori arguments provides a very strong case indeed that the normative principles of rationality serve a prescriptive/self-critical function, but cannot be employed for the purposes of explanation or prediction. We saw earlier that it is easy to see how it would follow that explanations at the personal-level can be fully understood without knowing any facts at the subpersonal level (that is, the first strand of the notion of autonomy) if the inferential principles presupposed by and tacitly applied in psychological prediction and explanation, as well as practical decision-making, were the familiar principles of the logic of consistency and the logic of truth. But we now know that these are not the relevant inferential principles. Nonetheless, we will not have a full understanding of explanations at the personal level until we know what the inferential principles governing personal-level explanation actually are.

But it should now be apparent why this threatens the putative autonomy of personal-level explanation. Understanding the inferential principles which govern personal-level explanation is understanding a subpersonal fact. Personal-level explanation is governed by the particular set of inferential principles that happen to be psychologically real – just as language comprehension is governed by the particular set of syntactic principles that happen to be psychologically real. As the experimental work on reasoning shows, there are many ways to interpret the behaviour of putatively rational agents. Take the selection task, for example. We can interpret the experimental subjects as confusing conditionals with biconditionals, but correctly applying the principles governing the biconditionals.⁸ We can view them as misapplying the principles governing conditionals. We can

⁷ Schick takes a less drastic view.

⁸ If the statement ‘if there's a circle on the left then there's a circle on the right’ is read as ‘there’s a circle on the left if and only if there’s a circle on the right’, then it can be properly evaluated without turning card (d) over. By the same token, it will be necessary to turn over card (c).

view them as correctly applying some pragmatic reasoning schema - or as employing some principle for reducing information-theoretic uncertainty. There are all sorts of personal-level stories that can be told consistent with the experimental evidence. But they cannot obviously all be true. Only one can, and that is the one which appeals to principles that are actually included in the tacitly known “mental logic” that controls the unreflective instances of both practical and theoretical decision-making.⁹

§5 Skilled behaviour and the autonomy of personal-level explanation

Recall that the claim that personal-level explanation is autonomous has two distinct strands:

- a). explanations at the personal-level can be fully understood without knowing any facts at the subpersonal level.
- b). subpersonal states will not feature in (horizontal) explanations of behaviour at the personal level, when that behaviour is being explained in the manner distinctive of personal-level explanation.

The previous section took issue with the first strand. The second strand is the target of this section.

To provide some background consider the following passage from Jennifer Hornsby. She is considering the proposal that common sense psychological explanations of behaviour need to be supplemented by subpersonal psychological facts about bodily movements. She starts by describing the reasoning that she discerns behind the proposal:

It is as if common-sense psychology had a hidden complexity that the theoretical psychologist could uncover experimentally; as if the superimposition of the picture of the person on the picture of the brain could reveal a sort of complexity in the picture of the person which ordinarily goes unheeded. (Hornsby 1986, 106)

Here’s what she thinks is wrong with this superimposition:

If common-sense psychology has no concern with how exactly we move our fingers when we turn on lights (say), then this is because we do not have to try to move our fingers in the exact way in which we actually move them in order to turn on a light when we want to. But where the details of bodily movements are not within common-sense psychology’s province, how can that which bears on the details have a bearing on common-sense psychological states. (Hornsby 1986, 106-7)

⁹ There is little if any reason to think that similar principles govern reflective and unreflective decision-making. In fact, on the plausible assumption that unreflective decision-making is phylogenetically much more primitive than reflective decision-making there is every reason to think that the principles governing reflective and unreflective decision-making must be dissimilar.

The point here, I assume, is that the light coming on does not depend upon my making precisely this set of (subpersonally specifiable) bodily movements. On the contrary, there is a sizeable class of bodily movements which are interestingly picked out at the personal level by the fact that they all have in common the fact that they will satisfactorily enable me to switch on the light when I so wish. Any one of them would have been adequate, and so there's no need to go into details about the particular bodily movements involved.

Let me put this another way. It has been argued by many philosophers, most famously Davidson, that personal-level explanations are causal, but not governed by causal laws. Hornsby and other defenders of the autonomy of personal-level explanation endorse this assumption. There is a *prima facie* puzzle, however, as to how an explanation can be causal without being governed by causal laws. Solutions to this puzzle fall into two broad strategies (Ruben 1994). On one strategy, endorsed by Davidson himself, explanations at the personal level are systematically connected with explanations at the microphysical level, with the latter explanations providing the essential causal laws. There are familiar difficulties with this strategy, deriving from the difficulty of specifying these systematic connections in a way that is neither empty nor opens the door to a form of reductionism (Kim 1993 Pt II). Some philosophers have correspondingly been attracted to the second strategy, trying to develop an account of causal efficacy that employs only concepts available at the personal level. Proponents of the autonomy of personal-level explanation will naturally be attracted to this second strategy.

Counterfactuals are an essential part of the second strategy. They do the work in explaining why personal-level explanations are explanatory that is done by causal laws outside the realm of the personal level (Schiffer 1991, Ruben 1994). I have taken the following proposal from Schiffer as a template for the general form that any such account will take¹⁰:

The F causally explains the G iff:
 (a) the F caused the G
 (b) if there had been another token event c* in place of the F, which failed to be an F, c* would not have caused the G

¹⁰ Ruben makes a series of complex refinements to the proposal in his excellent article, but none of them affects the point I want to make

It is easy to see how, on an account like this, Hornsby is quite right that explaining why the light coming on does not involve reference to my making precisely this set of (subpersonally specifiable) bodily movements. Taking the F to be the particular set of bodily movements which actually resulted in my switching on the light, and G to be the light coming on, it is clear that the second component is not satisfied, because many different sets of bodily movements could equally have resulted in the light coming on. If, on the other hand, we take the F to be a flicking of the light switch, or some such member of an intentionally characterised family of actions, both clauses are satisfied.

Things get slightly more complicated, however, when we look at an example of skilled behaviour more sophisticated than switching on a light. Suppose that I am playing tennis and lose because my opponent plays a brilliant backhand drop volley when it is match point. What is the personal explanation of this? Is it that he is more determined than me, that he wants to win more, that he has trained harder, that he is more skilful? All these things are no doubt true. And they are all part of the explanation of why he beat me. But his skill, training and determination are all perfectly compatible with his losing the game. They don't explain why he played a backhand drop volley at that particular moment. And yet that's why he beat me – because of that particular shot at that particular time.

There's a very real problem here. No explanation proceeding purely in terms of personal level events will be sufficiently fine-grained to explain why that shot was played then – to explain why my opponent extended his racket precisely that distance at precisely that angle. On the other hand, this seems to be a paradigm of a personal-level event. After all, he won the game and we can't explain why he won the game without explaining why he played that shot then. For that we need to advert to details of the integration of vision and action at the subpersonal level. We need to explain how a perceptual registering of a ball arriving at a particular angle, together with a perceptual registering of where the opponent is on the court, are translated into particular configurations of muscular movements. These are, of course, subpersonal facts, and they will feature in horizontal personal-level explanations in precisely the way that the second strand of the autonomy thesis rules out.

The disanalogy between this case and the earlier neuropsychological example is important. The neuropsychological case showed that subpersonal states could feature in horizontal explanations of personal-level states, but such personal-level explanations do not qualify as full personal-level explanations because they are not governed by considerations of rationality. In the case of skilled behaviour, however, it seems clear that we are dealing with full personal-level explanations. My opponent won the game because he played that shot, but his playing that shot was equally a matter of his intending to aim a shot just out of my reach, of his desire to win the game and so forth. Correspondingly his shot is assessable according to standards of appropriateness and rationality in the light. The subpersonal details are an essential component of a broader explanation that remains firmly personal-level according to the criteria discussed earlier.

The important point is that this need to bring in subpersonal details is very much in contrast to 'ordinary' cases of intentional action, like Hornsby's example of switching on a light. As I mentioned earlier, my switching on the light does not depend upon my making precisely this set of subpersonally specifiable bodily movements. I could have made a significantly different set of movements and it would still have been the case that I was switching on the light because I wanted to. The following conditional, therefore, is false:

(C) If I had not made that particular set of bodily movements then the light would not have come on.

And the falsity of this conditional explains why an explanation of why the light comes on should not include precise details of bodily movements. But things are rather different when highly skilled behaviour is involved, as it is in explanations of why I lost the game of tennis. Here my losing the game is a result of my opponent's making precisely the movement that he did make in that given situation. It seems to me that the corresponding conditional is true in the case of highly skilled behaviour:

(C') If my opponent had not made that particular set of bodily movements then I would not have lost the game of tennis.

Recall the counterfactual constraint that I mentioned earlier. Here it is again:

The F causally explains the G iff:

- a) the F caused the G
- b) if there had been another token event c^* in place of the F, which failed to be an F, c^* would not have caused the G

When the relevant F is taken to be the particular set of bodily movements made by my opponent, the truth of conditional (2) just mentioned ensures the satisfaction of the second component. No specification of the backhand shot in personal-level terms will satisfy the (b)-clause, and hence be genuinely explanatory.

Note, moreover, that it is not possible to deal with this in the same way as the blindsight example was dealt with in §2. There we saw an example of a personal-level state that could only be horizontally explained by including certain subpersonal states as its direct antecedents. It proved possible to reconcile this with the autonomy thesis by holding that subpersonal states can only feature in explanations of personal-level behaviour when that behaviour is not being explained in the manner distinctive of personal-level explanation, that is, in a way that is subject to assessment in accordance with criteria of rationality.

Of course, it is open to the defender of autonomy to reject the model of personal-level causal explanation that I have employed – most simply by developing an alternative model, or more drastically by rejecting the idea that personal-level explanation is causal. The drastic strategy is obviously doomed, however, and any model likely to avoid the difficulties with skilled behaviour I have identified will have to eschew the use of counterfactuals. I have no idea what form such a model could take – unless it involves a shift to the first strategy, which poses its own reductionist threat to the autonomy of personal-level explanation.

§6 Conclusion

This paper has tried to make some dents in the view that there is a distinct and autonomous level of personal-level psychological explanation. Experimental work on reasoning habits suggests that the rationality constraints constitutive of personal-level explanation can only be fully understood if certain subpersonal facts about the hard-wired inferential principles governing unreflective explanation and prediction are understood. And the analysis of skilled behaviour suggests that subpersonal states will have to feature in the horizontal explanation of a central class of personal-level states. Of course, these are just dents, and it will probably need quite a few more dents before the autonomy thesis is as damaged as I would like it to be. Let me conclude, though, by offering one

very general reason why I think that the autonomy thesis is philosophically of the highest significance.

The reason is this. Belief in the autonomy of personal level explanation can make eliminativism about personal-level psychological states very appealing. A standard argument for eliminativism is that personal-level psychology fails to reduce to more basic subpersonal levels of explanation. We won't, so it is argued, be able to find subpersonal relations that map all and only the rational relations between personal-level states, and so we cannot accept the legitimacy of either personal-level states or the explanations within which they feature. But it is only if one thinks that personal-level states, if they exist at all, must feature in an autonomous domain of explanation that one will be tempted to expect any sort of reduction of explanation involving personal-level states to explanations not involving such states. A second frequently cited argument for eliminativism is that personal-level psychological explanation is inadequate, because it cannot deal with, *inter alia*, skilled behaviour, neuropsychological cases and apparently irrational behaviour. The solution to this, it seems to me, is not to eliminate personal level psychological states, but to make these phenomena more tractable by eliminating the autonomy thesis. Personal-level psychological explanation cannot deal with, for example, skilled behaviour while confining itself to the explanatory resources of the personal level, but that explanation involving personal-level states should confine itself in that way follows only if we accept the autonomy thesis. Without the autonomy thesis the eliminativist argument loses its bite.