Contemporary Pragmatism Editions Rodopi
Vol. 1, No. 2 (December 2004), 1-66

Naturalizing Idealizations: Pragmatism and the Interpretivist Strategy

Bjørn Ramberg

[80 to 100 word Abstract]

Let us say, with Quine, Davidson and Dennett among others, that a person's language and psychological attitudes have their identities fixed with the theories generated by an idealized interpreter of that person (Quine 1960; Davidson 1984, 1986a, 1986b, 1989a, 1989b, 1990a; Dennett 1978, 1987a, 1991a). A reason for saying this is that it will help us see how the capacities to entertain attitudes and to communicate linguistically can be natural capacities, capacities we may happily attribute to creatures who fall squarely within the scope of evolutionary biology. This, at any rate, is Rorty's principal reason. The interpretivist strategy permits us, Rorty suggests, to give an account of persons which introduces

> no breaks in the hierarchy of increasingly complex adjustments to novel stimulation — the hierarchy which has amoebae adjusting themselves to changed water temperature at the bottom, bees dancing and chess players check-mating in the middle, and political revolutions at the top. (Rorty 1991b, 109)

How does it do this? I will develop an answer emphasizing the naturalistic motivations of the interpretivist strategy, an answer that is also intended to draw out and situate some of the commitments underpinning the view of philosophy that Rorty has worked out over the last thirty five years (1967, 1979, 1982, 1989, 1991a, 1998b, 1999).[1] While this combination of constructive polemic and metaphilosophical commentary makes for a long paper, the view of Rorty's pragmatic philosophy that I want in this way to make vivid can be stated briefly. Rorty's thought represents a dialectical

transformation of naturalism. As he brings naturalism to bear fully on the project of philosophical reflection itself, Rorty finds himself fundamentally changing the requirements we impose upon our thinking whenever we seek to assume a naturalistic philosophical stance toward some subject matter.[2] To appreciate the naturalizing capacity of the interpretivist strategy is to understand how Rorty's naturalistic critique of philosophy alters the nature of naturalism itself.

There is an ulterior purpose behind this indirect approach to Rorty's pragmatism. Rorty laments the tendencies toward "decadent scholasticism" of contemporary professionalized philosophy. (Rorty 1993a, 1995b) The response he advocates has been, on the whole, the complete overthrow of the vocabularies in which much contemporary professionalized philosophy is carried out. The vocabularies of epistemology, of content-oriented philosophies of mind and language, the vocabulary of meta-ethics, these are all, to Rorty's ear, corrupt beyond redemption. However, it seems to me that Rorty would give up nothing of substantive importance, and indeed be better placed to reclaim some useful fortifications from which to combat the decadent scholasticism that he often astutely exposes, if he were less revolutionary inclined toward the historically developed vessels within which much philosophy is presently conducted. Rorty could afford, for instance, to be less reticent than he has so far been about invoking notions like 'rationality' and 'norms of reason'. Now, the differences in the respective depictions of the concept of rationality in Rorty (1989) and Rorty (1999) indicate movement in the direction I would urge; the former is dismissive (e.g. 1989, 47), whereas the latter cautiously suggests that 'rationality' is a reconstructible notion that can be made to do useful work. I think the point generalizes to pretty much all the terms that in philosophical contexts arouse Rorty's deep suspicions; "knowledge," "truth," "content," and the like. As I see it, the best remedy for decadent philosophy is to conduct an aggressive campaign of pragmatizing reappropriation of just those terms that traditionally have been employed to express ahistoricist and essentialist conceptions of reflection. If this essay were successful it would lend illustrative support to this claim, and help pave the way for an invigorated, assertive, historically self-conscious brand of pragmatist philosophical reflection that I like to think of as *Revisionist Rortyanism*.[3]

1. Pragmatic Redescription *versus* Philosophy of Mind

A distinctive feature of the interpretivist strategy as it has been developed after Quine (1960), is that it aims for naturalization without taking the route through nomological or conceptual reduction. Where some see only three alternatives — some form of reduction, outright elimination, or a retreat to dualism — the

post-Quinean interpretivist claims to mark out a fourth possibility.

The coherence of this possibility can certainly be doubted. Fodor, for example, persistently argues that the interpretivist's theory of the attitudes, with its inevitably ensuing holistic individuation, is really a coy form of eliminativism.[4] For Fodor, the honest position to take, if you must be an interpretivist about the attitudes, is that of Quine (1960).[5] Kim, to take another example, has no less persistently argued that a commitment to monism, to physical predicates as the proper constituents of basic laws, and to the reality of the attitudes, makes the reduction of the mental to the physical inexorable.[6] Such arguments are typically rooted in firmly intuited constraints on what it really is to consider something as real, intuitions that yield the metaphysical conviction that naturalism and reductivism (or eliminativism) are inseparable.[7]

Pragmatists will treat such ontological intuitions as symptoms to be examined, not as foundation for argument. They will see them principally as expressions of commitments to particular *vocabularies*.[8] The attempt to settle what the reifications of a vocabulary *really are*, in terms of some other, ontologically legitimizing vocabulary, is itself at odds with a naturalistic view of thought. Pragmatists do not want to say that the mental is really something physical or material. Nor, though, do they want to say that, really, it is something non-material or non-physical. Naturalistic pragmatists are proposing ways to describe ourselves as thinkers and agents that make the philosophical contrast between mind and matter seem to be without any particular ontological point. Perhaps one might signal this sort of attitude by calling oneself a non-reductive physicalist (Rorty 1987). My strong suspicion, however, is that it is not very helpful to try to spell out the antidualistic commitments of a pragmatized naturalism in terms of its relation to physicalism. 'Physicalism' — in all its varieties with their attendant conceptual distinctions — is burdened with the connotations of a dichotomous folk-ontology, one that has been hypostatized in the terms of art of the kind of philosophical vocabulary to which naturalistic pragmatists are busy working up alternatives.

Indeed, our notion of mind and the vocabulary in which it is embedded well illustrates how philosophical analysis and 'intuition', providing mutual support and reinforcement, can entrench a particular set of problems and make them appear mandatory. Unfortunately, though, it could also be taken to bear out the anti-pragmatist point that "mere coherence" is not enough; we need a touch-stone against which to test the truth of even the most reflectively equilibrated beliefs. If not a priori reflection, then empirical science may provide just such a touch-stone — so long as we believe that science can aim to articulate a description of the world warranted by criteria that are demonstrably truth-indicative (e.g. Haack 1993). Demonstrably truth-indicative criteria, we realize, are ones that normative epistemology will show

we have good reason to believe point us toward the way the world is, in the way we have good reason to believe that a compass will point us toward the Magnetic North Pole. If we fail to perceive the conceptual connection between the very idea of justification — or assertoric warrant — and a distinct truth-norm (e.g. Haack 1995, Wright 1992), a connection which allows us to draw a distinction between genuine, objective warrant and mere assertion-games, then we are stuck with parochial coherence as our only measure. The result is a kind of idealism without the innocence, a jaded ironism with no recourse to rational means of settling theoretical (or practical) conflict. A charge against the pragmatic view I defend is precisely that this is just where it leaves us (Haack 1995).[9]

What the pragmatist suggests, however, is that this very construal of enquiry and of warrant and of truth is forced on us by the assumptions embedded in an entrenched vocabulary of mind. This vocabulary leaves much of philosophy preoccupied with conceptual problems the various proposed solutions to which generally float quite free of the practical and theoretical problems that engage us as the 21st Century gets under way. The pragmatic philosopher treats such conceptual problems as points of leverage for vocabulary shifts. Tracing questions posed in the vocabulary of mind back to the assumptions that make them appear compelling, pragmatic philosophy is self-consciously historicist. This is not, it is important to note, to *reduce* philosophy to the telling of the history of philosophy. It is to oppose a conception of philosophy that treats the history of the subject as a more or less valuable heuristic aid to reflection. The key historicizing move of the pragmatist is to temporalize meaning, and so to treat content in socio-genetic terms. This move is what makes advancement in philosophical understanding inseparable from the telling and retelling of reconstructive histories of the problems we are trying to understand.[10] The pragmatist will, accordingly, offer genealogies of philosophical problems, genealogies which aim to redescribe our philosophical urges and inclinations in such a way that we can extricate those theoretical aims we may want to stand by from what has appeared to be mandatory frameworks for their articulation.[11]

Now, the interpretive strategy plays a crucial role in this effort, because, pragmatists hopefully believe, it will allow us to precipitate out a vocabulary of agents and thinkers from the vocabulary structured around that pair of intimate antonyms, ‘mind’ and ‘matter’. Pragmatists hope that this will, eventually, undercut the governing intuition of Philosophy of Mind, the conviction that the kinds we capture with psychological ascriptions just could not in themselves, at least not straightforwardly, be natural states of natural creatures. Pragmatists do not believe that our practice of psychological ascriptions leads us inexorably to the mind-body problem. Rather, they see in ‘mind’ the vestiges of ‘soul’, and hypothesize that the real problem is actually our deeply-rooted

attachment to this ancestral notion — and our concomitant commitment to the idea that the relation of 'mind' to its conceptual counterpart is a central philosophical difficulty. It is this attachment that makes it appear *prima facie* mysterious how the vocabulary by which we are able to treat some things as agents could capture a way that some natural creatures (and, perhaps, artificial systems) are in the world. What the pragmatist polemic takes aim at, then, is this attachment, this deep-rooted commitment. This, for the pragmatist, is what philosophically motivates the interpretivist strategy.

The pragmatist is not claiming to solve the mind-body problem, nor to dissolve it. Nor is the problem being diagnosed as illusory, as a product of some form of conceptual confusion, linguistic mistake or general lack of semantic alertness. The pragmatist takes the mind-body problem to be real, but transient. It is a problem we will come to see as idle once we have developed better ways of conceiving ourselves and our relations to our surroundings, once we have developed, that is, better vocabularies. These vocabularies will be better in the specific sense that they will enable us to treat certain items as agents *without* sticking us with dichotomous schemes of fundamental ontological kinds, the kind of kinds whose relation one to the other cannot but become immediately problematic. The interpretivist strategy is attractive because it holds out the promise of just this kind of improvement in our conception of the capacities that make us persons.

Dennett may be used to illustrate the approach I have in mind. Contrast Dennett's attitude to the attitudes with his attitude to consciousness; to put it crudely, the attitudes survive (1987a, 1991a) while consciousness must go (1991b). Why is this? For Dennett, a principal tool in the campaign against intuitions of mind and the reifications that philosophy spins from it is natural science. Then, has neuroscience discovered the underlying states that we might plausibly take belief-talk really to refer to? Have neuroscientists determined that, really, there is no such thing as consciousness? This, we know, is hardly Dennett's point (cf. 1981c, 93; 1991b, part 3, appendix A). Dennett is motivated by the diagnosis that the folk-notion of consciousness keeps us wedded to a set of interwoven descriptions of mind and self that inhibit our susceptibility to the naturalizing influence of science on our self-image. This set of descriptions is what we gesture at with the notion of the subjective. The sense that the notion of the subjective is a rich and *bona fide* mine of philosophical problems and insights is an explicit target of Dennett's seditious account of mind.[12] Dennett's view is that the linguistic practices in which our notion of consciousness is embedded (the vocabulary from which the philosophical invention 'qualia' takes its intuitive power — see Dennett 1988, 1991b), are practices we would do well, if we want to naturalize our conception of ourselves, to alter. But this, any pragmatist knows, we can do only in so far as we are able make satisfying alternative descriptions available.

As I read Dennett (1991b), his ambitious account of consciousness exploits the tensions at the interface of science and common sense in order to alter our sense of what is obvious about awareness, of what are the base-line home truths of our experience of being what we are. Seeking to instil changes in the linguistic habits that make up the vocabulary of the phenomenology of experience, Dennett hopes to make descriptions arising from natural science mesh more easily with our sense of what matters about us as persons.[13] The vocabulary of the attitudes, by contrast, is in Dennett's view redeemed by the interpretivist strategy. Why? Because, like Rorty, he takes it that this account shows how it is that attitude ascriptions can be useful to the point of indispensability to natural creatures like us.[14]

I am about to offer a version of the interpretivist strategy that will enable me to make explicit its intimate connection with a pragmatist conception of rationality and of philosophy. I will be stressing two closely related aspects of this intimate connection. Pragmatism serves interpretivism, in so far as an effective defence of the interpretivist strategy against common objections will appeal to a pragmatic conception of rationality. Interpretivism serves pragmatism, in so far as the strategy becomes, in the context of the conception, a tool for naturalization. A crucial question, then, is this. How can we tell whether the interpretivist strategy will actually do the job the pragmatist wants it to do? This question is not yet the question of how the strategy can be an instrument of naturalization; the latter issue is metaphilosophical, and turns on our assessment of the conceptual resources that interpretivists employ in formulating the strategy. The prior question asks simply how we may assess the interpretivist's claim to be giving an account of content, whether that account is regarded as naturalistic or not: does the strategy produces a convincing account of what we, the folk, recognize — or, better, can be brought to recognize — as key features of the lives of persons?

The states with which the interpretivist is concerned — the states we invoke when describing creatures as agents and thinkers — are anchored in our attributive practices of run-of-the-mill interpretation and psychological explanation, and these practices provide the measure of plausibility. Consequently, an important kind of argument against interpretivists is one which drives in a wedge between what a given version of the approach says about the individuation and attribution of psychological states and the kinds of relations these may stand in to each other, on the one hand, and the individuative and attributive practices embedded in our ordinary use of folk-psychological terms, on the other. Much of what follows will address this sort of argument. In particular, I will examine claims to the effect that the interpretivist strategy is incompatible with the degree to which and the manner in which the irrational and the non-rational enter into our accounts of persons.

2. The Interpretive Strategy and Two Reasons Why It Seems So Implausible

Consider the ideal interpreter, IDA, let us say. IDA is a thoroughly theoretical being, whose essence it is to implement a specified methodology of interpretation. In so doing, IDA is purported to provide a model for a certain kind of ability or competence that we actual interpreters appear to have. However, the methodology in question has precious little to do with the actual methods of field linguists or translation-manual constructors. The point of this methodology is to make manifest a way to view the sorts of concepts that we apply essentially in descriptions of agents and thinkers. The relation of the methodology of ideal interpretation to the actual capacities of actual interpreters is captured in the following question: could we, if we possessed the knowledge about some person expressed in IDA's theory, plausibly be said to understand that person? The issue here is how the interpretivist's proposed account of the nature and point of psychological attitudes and linguistic meaning, as expressed in the constraints on ideal interpretation, is tested against folk-psychological practice. In so far as IDA appears capable of coming up in a given case with attributions and ascriptions that harmonize with those of actual interpreters, this provides support for the view of the nature and point of these attributions and ascriptions that the interpretive strategy is devised to make explicit.

The interpretive strategy is intended, then, to tell us something about how we should think about what it is we are doing when we engage in psychological attribution and semantical ascription. It should be noted that on this construal of its theoretical point, IDA's methodology has no particular normative implications at all, even implicitly, for us actual interpreters, eager, as we ever are, to improve our understanding of our fellows. It may turn out for some characterization of IDA that the conclusions drawn on the basis of the evidence we allow end up diverging from what we should want, intuitively, to say about the subject of the interpretation. In that event, and to that extent, the relevant specification of ideal interpretation would lose its point. It would cease to play a useful role in our attempt to illuminate the vocabulary of thought and action.[15]

IDA will be idealized in several ways, of which the following are among the more conspicuous. For one thing, IDA will be cognitively idealized; IDA's ability to construct and modify explicit theories in light of evidence, and to assess their relative empirical merit, their adequacy to the evidence, is unencumbered by the contingent characteristics that keep actual theorizers from contemplating in principle available alternatives. Further, the evidential base for IDA's theorizing is one no actual interpreter could ever rely on. Not only will IDA observe everything subjects of interpretation do, including, of

course, the noises they make, and the environing conditions of all this activity; IDA will also have access to the behavioural dispositions of interpreted subjects. That is to say, for purposes of theory-construction IDA is assumed to be able to appeal to the truth-values of counter-factual conditionals of a kind that actual interpreters would have to treat as untested predictions. Finally, IDA is ideal in being without preconceptions, both as to the semantic value of particular vocables, movements or inscriptions produced by the subjects, and as to the particular details of the subjects' intentional relations to the world.[16] This last point we might put by saying that IDA has no initial view of the particulars of the pattern of truth-preferences that are distinctive of some arbitrary subject of interpretation.[17] What IDA must have, however, is a view of certain general features of any such pattern of preferences; IDA must operate with certain desiderata that any set of attitude-ascriptions should conform to. Otherwise, the idealized observational access to the subject's behavioural dispositions and their contexts would do no good, because nothing would constrain the inferences IDA may draw from that evidence. There would be nothing in particular that the "evidence" could be counted as evidence for, and so it would not be *evidence* at all.[18]

The central task for the interpretivist is to make explicit the empirical methodological constraints under which IDA is to deliver her specifications of meanings and attitudes. Specifically, the interpretivist must characterize those general features of truth-preference patterns that allow IDA to see observed events as evidence for particular theories of meaning and belief. This characterization is what displays the view of the vocabulary of thought and action that the interpretivist recommends. It must, on the one hand, serve the naturalizing motivation for the pragmatist's deployment of the interpretivist strategy, while securing, on the other, convincing results when put to the test by means of IDA. An initial characterization might be: IDA must structure her descriptions of the actual and possible events that serve as evidence in accordance with the pattern of reason. The suggestion here, familiar from the writings of Dennett (1987) and Davidson (1984), is that a subject's perspective on the world revealed by interpretation inevitably emerges as a rational one. The point of the suggestion is this. What it is to be a belief or other psychological attitude is to be a state in a network of states that allows us to see a significant segment of the behaviour of some creature as manifesting a rational orientation to its environment. According to this position, attitude attribution discloses *a point of view* on the world, the particular nature of which is traced by those ascriptions.[19] By the terms of this point of view, some subset of its occupier's causal transactions with her environment are seen to serve intelligible purposes. That intelligibility is what gives the attitude-scheme its value — to serve our predictive needs, as Dennett (1981a, 1981b, 1991a) emphasises, perhaps predictive interests of a particular sort, as

Davidson (1991a) hints. I will be revisiting the important connection between interest and content ascriptions at several points throughout the paper.

The essence of this view is that in so far as we are dealing with creatures (or machines, or what have you) qua agents, the better theory simply is the one that better rationalizes behaviour. Of the many theories that could be made to account for the evidence, *the optimal theory for IDA has the subject(s) less beset by irrationalities than do alternative theories*.[20] This is to assert an unrepentantly rationalistic version of the methodological constraint on ideal interpretation, one we might therefore label the Rationality Maxim. It will be important to keep in mind, as we assess objections to interpretivism, that RM has what we may call global scope. That is to say, IDA relies on RM to choose between candidates for total theories — or, in anticipation of a later distinction, for total accounts. Just because it constrains theory-choice holistically, RM governs the interpretation of any particular utterance or movement only in an indirect, mediated way. The kind of rationality-judgements we will require IDA to be guided by are going to be over-all judgements of the global state of subjects captured or characterized by various candidate theories or accounts.

At any moment or stage of interpretation, then, RM constrains the simultaneous attribution of the entire gamut of intentional attitudes: according to it, IDA will endeavour to make a subject, at a time, believe what is true and cherish what is good, dread the terrible and yearn for the lovely (cf. Davidson 1970, 222). The demand imposed by RM is not only a demand for consistency among a subject's beliefs and attitudes, and for coherence among the subject's means of describing the world. Rationalizing a person by RM, IDA will seek to have the subject prefer true the right sentences — that is, just those sentences which, as IDA interprets them, the subject *ought* to prefer — and to prefer them, moreover, *for the right reasons*. Aiming for global rationality will not single out a class of attitudes, such as beliefs regarding matters of fact, rather than, say, matters of method or matters of value. There is no fact-value gap nor truth-method gap in ideal interpretation. And since noises are speech only when situated in a general context of agency, having subjects prefer true the right sentences IDA must also have them do the right thing. In short: applying RM, IDA insists, as far as possible, on her subject's cognitive and moral perfection.[21]

A natural and frequent objection is that this way of characterizing the methodology of the interpreter must be wrong, since people patently are not impeccably rational, as this injunction to IDA appears to be presupposing. We need, this line of criticism has it, to let various well-known impediments to such perfection come to expression in IDA's methodology. Different versions of this challenge have been directed at interpretivists since Grandy (1971) criticized Quine's (1960) deployment, in his account of radical translation, of Wilson's (1959) principle of charity. This kind of objection is directed at the

content of RM, and not (so it seems) at interpretivism per se. After all, interpretivism might fly with a more modest (or, as I shall say, with a weaker) methodological principle at its core. In Sections III and IV below, I attempt to defang a careful statement of this position — which, paraphrasing Grandy, I will call the humanitarian position — as offered by Føllesdal (1982a). After elaborating the humanitarian view, I argue for two distinct claims. Firstly (Section III), interpretivism cannot depend on a psychologically qualified maxim with a weaker rationality-demand than that of RM and still serve to explicate the point and function of our vocabulary of psychological and semantic ascriptions. If we were to adopt the humanitarian position we in fact give up on interpretivism as a naturalizing strategy. In making this point, I will elaborate and make use of the Rortyan notion of a vocabulary. Secondly (Section IV), contrary to arguments underwriting humanitarian claims, RM-based ideal interpretation can indeed be squared with the considered preferences of actual interpreters among competing ascriptions. Accommodating RM to such preferences yields, I argue, a view of meaning which is radically contextualist, and attractive to pragmatists on independent grounds. In Section V, I emphasize the distinctively Nietzschean nature of this notion of the contextuality of content by contrasting it with the 'locality-of-meaning' thesis advanced by Bilgrami (1992).

Even if humanitarian objections can be met or deflected, a further worry is highly salient in the present context. It is hard to see how an unabashed appeal to "the norms of reason" of the sort issued by way of RM to our ideal interpreter could sustain any serious naturalistic ambition.[22] I hope to make this less difficult by elaborating, first, the pragmatic nature of the conception of reason that informs the interpretive strategy; and second, a pragmatic conception of what naturalization demands. My point of departure (in Section VI) is an objection directed explicitly at the interpretive strategy itself. However we formulate the maxim that guides IDA — whether charitably (RM) or humanistically — we still secure some degree or other of rationality for psyche-endowed creatures by philosophical fiat. But surely, the objection goes, how good we are at thinking, and how well we act, must be empirical questions. Fodor (1987) has claimed this, though with more passion than argument. Stich (1981, 1990) has pressed the point by drawing on the plausibility of certain kinds of research programmes in cognitive psychology. I argue that the interpretivist begs no question of empirical import with respect to the quality of our cognitive capacities. Continuing to use Stich as a foil, I go on to suggest that doubts about the interpretivist's commitment to naturalism arise in part from a conception of rationality — and thus of the claims the interpretivist makes about rationality — that interpretivists need not, and should not, buy into. Then, in Section VII, I claim that the antireductivism of the interpretive strategy is incompatible with naturalism only on certain

metaphysical assumptions. These assumptions are directly challenged by Rortyan pragmatism. For the pragmatist, irreducibility emerges not as a reflection of a metaphysical gap, but as an ontologically innocuous reflection of the divergent human interests that vocabularies serve. What needs naturalizing, I suggest, is not this or that descriptive practice, but philosophy. In the concluding section (VIII), I draw some consequences of my Nietzschean version of interpretivism for our conception of philosophical reflection. The conclusion I offer runs against the impression that Rorty sometimes creates; we can stake a claim for the irreducible distinctiveness of reason and for philosophical reflection without betraying Rorty's pragmatic naturalizing of philosophy.

3. Why Pragmatists Are Not Humanitarians

Ideal interpretation minimizes the unintelligibility of the agent. Thus stating the obvious, this flexible formulation may seem to suggest a final convergence of various proposed versions of principles of humanity and charity (see e.g. Evnine 1991).[23] It seems compatible, for example, with the much fuller version of the empirical constraint on interpretation that Føllesdal (1982a) has given. Føllesdal provides a clear, characteristically circumspect examination of the status of rationality as a constraint on interpretation, and suggests a formulation that well captures the point of humanistic qualifications of charity. When ascribing attitudes to persons, Føllesdal suggests,

> on the basis of observation of what [they do and say], do not try to maximize ... rationality or ... agreement with yourself, but use all your knowledge about how beliefs and attitudes are formed under the influence of causal factors, reflection, and so forth, and in particular your knowledge about [their] past experience, [their] various personality traits, such as credulity, alertness, reflectiveness etc. Ascribe to [them] the beliefs and attitudes you would expect [them] to have on the basis of this whole theory of [persons] in general and [individuals] in particular. (1982a, 315–316)

It might look like Føllesdal's catalogue of explanatory avenues could serve as a specification of the general demand to minimize unintelligibility. However, this apparent convergence of humanity and charity is illusory. The issue between rationalists and humanitarians remains sharp.

Rationalists (I stipulate) endorse RM; what counts as justifying the attribution of a content and an attitude toward it is to assign to that state a location in a pattern of intentional states in just such a way that the pattern as a whole minimizes the irrationality of the agent. That, says the rationalist, just is

what it is to capture the first-person point of view of some agent. It may certainly happen in a particular case that imputation of a bad attitude (a false belief, a strange desire) represents a better interpretation than does a good one, but that is because the preservation of local truth or goodness sometimes detracts from the global preservation of virtue for which IDA, according to the rationalist, must strive.

Humanitarians, by contrast, do not think that we make an agent's perspective on the world manifest by minimizing the irrationality of the subject. They want to introduce constraints on interpretation that modify or qualify the injunction to minimize irrationality; they believe that the justification of an attribution may be secured by reference to psycho- and socio-biographies, or to physiological theories and histories. With Føllesdal, they think that empirical theories of how actual human beings typically (mis)perceive, (fail to) reason, and (inefficiently) act — theories embodying causal generalizations that may be invoked in support of non-rationalizing attributions — can serve to make agents intelligible.

In taking this view Føllesdal stresses that theories of persons must be based on the recognition of "rationality as a second-order disposition" (1982a, 316), and so he endorses the view that action explanation is constitutively related to rationality. As he says,

> just as our theory of explanation of action must have room for ... deviant phenomena, so on the other hand the classification of something as deviant, as rationalization, as repression, sublimation, etc., is possible only on the basis of such a theory of how actions should be explained. (1982a, 310)

Nevertheless, Føllesdal takes this dependence of causal explanations on rationalization to be compatible with the view that explanations that do not serve over-all rationalization, but are grounded in empirical psychological theories, may be invoked in justification of an interpretation.

The intuitions underlying the humanitarian view of ideal interpretation seem compelling. Since we, as a matter of undeniable empirical fact, fall far short of our ideals of wisdom, circumspection, integrity, insight, and so on, this ought to be reflected methodologically in our conception of ideal interpretation. Granted, interpretation of persons must avoid that admittedly ill-defined limit where we have too much causal explanation and insufficient rationalizing going on to speak coherently of psychological states, but within this boundary causal constraints are not only possible, they are virtually self-evidently required. What we demand of a theory assigning psychological states and semantic values is that it captures an agent's perspective of the world. Since much thought and action is governed by irrational and non-rational

influences, a methodology of interpretation that construes us as though we were perfectly rational is less likely to produce the right theories than is one that explicitly takes our common short-comings into account. When we try to articulate some agent's point of view on the world, generalizations that bear on the nature and formation of that point of view are clearly relevant. They must therefore be built into the methodology of the interpreter, modifying the assumption of rationality.

Nothing in this picture seems explicitly to challenge the idea that the items assigned by the theory are in fact individuated by the theory. The humanitarian seems entitled to the claim that he is offering a version of the interpretive strategy. Why should we reject it? If we assign IDA a principle weaker than RM, such as Føllesdal's, we will foil the interpretivist's aspirations to offer an account of agency that is at once both non-reductive and naturalistic. The remainder of the present section is my attempt to make good on this claim.

The naturalizing potential of the interpretivist strategy rests in significant part on what Davidson calls "a bland monism." (Davidson 1970, 214) It is monistic, because it denies the dualist's thought that there are two ontological kinds; mental and physical. It is bland in a somewhat peculiar sense; it also denies the reductivist or eliminativist thought that there is *one* ontological kind of a sort to which our various ways of talking may stand in questionable relationship. The pragmatist thus takes the lesson of Davidson's (1970) argument for anomalous monism to be that we need not worry about the ontological priority of kinds of description, but only about their relative utility for specific purposes. Indeed, the naturalistic pragmatist encourages us to retreat altogether from ontology, advocating a view of language that simply leaves no room for it; the world causes our noises to mean what they do — by way of the complicated patterns of similarity-judgements that we endlessly interacting noise-makers are disposed to produce.[24] Reference, on this view, comes dirt cheap; a greater or lesser capacity for connecting us with what is really out there will not be what distinguishes one descriptive practice from another. We may, I suppose, still think of philosophical reflection as an attempt to illuminate what there is; but this cannot be construed as a matter of gauging the relative referential success of various descriptive practices. It becomes, rather, a matter of providing characterizations of the interests we have in referring to items of this or that sort.[25]

It is with respect to differences of such descriptive interests that we distinguish *vocabularies*.[26] When I insist on distancing the notion of a vocabulary from the concept of translation, it is not because, at least not just because, I happen to have fastened on a certain kind of use of 'vocabulary' as paradigmatic. Whatever other senses we can plausibly give the notion of a vocabulary, the one that I characterize, which has nothing in particular to do

with translation, is in any case central for many of Rorty's purposes. Brandom (2000) is absolutely right to suggest that for Rorty, a principal virtue of the 'vocabulary' vocabulary (as Brandom dubs it), is that it provides a way of designating discursive bodies that completely incorporates Quine's dissolution of any principled distinction between semantical and empirical commitments, as well as Davidson's devastation of the thought that the idea of a conceptual scheme is a philosophically interesting or fruitful one. What motivates Rorty's use of the concept of a vocabulary, is his thought that it may bring us closer to a philosophical vocabulary within which we may still the ontological urge, the urge that leads us to engage in projects of ontological legitimation. The concept serves this purpose precisely in so far as it allows us to pick out discursive structures in a manner that precludes any attempt to restore an ontologically potent form of the distinction between what we talk about and how we talk about it. I worry that to think of inter-vocabularic relations principally in terms of translation is to think in a way which may place all but the most self-consciously Quinean among us at odds with this purpose. The point of any vocabulary can be explicated only relative to the specific goals, needs and interests of its users or potential users. As is the case with other kinds of tools, what makes a vocabulary the particular vocabulary it is just is the particular manner in which it serves the needs and interests it serves. However, the relation between vocabularies and their uses differs from the relation between tools and their purposes in an important respect. Just as vocabularies cannot be individuated independently of the interests they serve, so these interests cannot be stated without employing the vocabulary. When we articulate the goals or purposes that give point to a vocabulary, then, we are offering an individuating characterization of that vocabulary, and making such a proposal is not distinct from providing a general description of the kinds of objects to which the vocabulary refers.[27]

When we claim to be characterizing a vocabulary, we thereby claim to be giving a *basic* account of some set of concepts. That is to say, we claim to be offering reasons for thinking that the interests we invoke, the concepts we analyze, and the manner of the analysis, all are linked in such a way that to use a different kind of concept would, *eo ipso*, be to serve different kinds of interests. Claiming to offer a basic account, in this sense, is not to rule out the possibility of there being — or coming to be — systematic conceptual relations between the vocabulary one thus specifies and other vocabularies. Rather, it is to insist that such conceptual relations will not provide a way for us to keep the interests as is and drop the concepts, in favour of those of some other vocabulary. If we decide to say, for example in a case where one explanatory paradigm replaces another in some area of enquiry, that this is actually what has happened, the conclusion to draw would be that our earlier conception of the vocabulary in question stood in need of revision; we had not fully grasped

what we were talking about. What we were working with was a pseudo-vocabulary; in so far as we had obtained a philosophical analysis of the vocabulary, it was one in which interest and concept turned out, in retrospect, not to be well matched.

Vocabularies are as enduring as interests are, which means that some will be highly transient, and others may be impossible for us to get by without. Like interests, they may be nested, contested, and individuated at cross-purposes. Further, we must not suppose that intellectual history will yield categorical diagnoses; emerging conceptual connections between vocabularies may lead to better, perhaps more comprehensive, accounts of vocabularies and interests, or they may indicate changes in interest, or themselves cause changes in interest. What may appear to one historian as the emergence of a better characterization of a vocabulary will to another appear as the abandoning of a set of goals in favour of another set. Such messiness tends to increase as historical distance decreases, approaching the chaotic at the limit constituted by the present.

Specifying interests, moreover, is itself an interest-governed enterprise — when we invoke vocabularies in our descriptions of social or intellectual evolution, no perspective is possible that is not laden with normative commitments. Similarly, any philosophical characterization of a vocabulary, staking a claim for the basic nature of some set of concepts, will involve a stipulative element. It will embody a proposal for conceiving of our interests in a certain way, a plea for seeing them that way and for assigning them a certain weight. The notion I am characterizing is essentially a hermeneutic one — vocabularies are never neutrally described, and they are never fully given.

Quite evidently, then, this heuristic notion of a vocabulary, pegged to the notion of interest, needs to be handled with some care. Nevertheless, it will presently serve a useful purpose. It provides exactly the right perspective on the interpretive strategy; this strategy is an attempt to make a case for a characterization of a vocabulary. As such it offers an account of a set of concepts, links the analysis of the concepts to certain interests, and holds the account thus offered to be a basic one, in the sense I have just characterized. Such a project cannot accommodate humanitarian modifications of ideal interpretation, as I will now argue.

What is distinctive, Davidson proposes, about "accounts of intentional behaviour" is that they "operate in a conceptual framework removed from the direct reach of physical law by describing both cause and effect, reason and action, as aspects of a portrait of a human agent." (1970, 225) Now, this is a claim that the interpretivist strategy is designed to preserve. As a constitutive account of a vocabulary of action, it aims to portray the rules governing the concepts of that vocabulary just so as to ensure the removal from law that Davidson speaks of. The interpretivist strategy does exactly this when it offers

us a view of these concepts whereby the very feature that gives them purchase on persons, free agents (as we redundantly say), at the same time renders them unsuitable as predicates of empirical law. A point of portraying concepts as governed holistically by rationality-considerations is to deprive those concepts of the particular kind of stability which empirical theorizing requires of its predicates; to the extent that some putative empirical generalization links psychological concepts in a way that is at odds with the norms governing them, to that extent the content of the generalization itself grows wobbly. This is just the feature of the concepts of the vocabulary that allows us to see ourselves and others as agents. What makes the vocabulary that Davidson aims to characterize the vocabulary it is, is its constitutive relation to agency.

Hence, when Davidson concludes that "[t]here cannot be tight connections between the realms [of the mental and the physical] if each is to retain allegiance to its proper source of evidence" (1970, 222), he is not just expressing a theoretical observation, he is expressing the very point of the rationality-constraint in ideal interpretation. That constraint is the centre-piece in a proposal which purports to make sense of agency by linking it constitutively to concepts that are identified exactly so as to cut across bodies of empirical, nomological generalization. The crucial point here is that this tight connection between particular interests and particular kinds of norms for application of concepts is what allows us to speak of a distinct vocabulary. It is only by virtue of its claim to offer an account of a distinct vocabulary, one incorporating the essential concepts of thought and action, that the interpretivist strategy can hope to provide a basic account of those concepts. This, in turn, is exactly what enables pragmatists to say that there is no further question of what intentional states are than what the interpretive strategy reveals.

It is this claim to be offering an account of a distinct vocabulary that the humanitarian version of the interpretivist strategy scuttles. On Føllesdal's view, holistic theories of persons hermeneutically balance causal psychological hypotheses and rationalizing interpretations in an attempt to account for all the behavioural evidence there is. Now, it is true that the balance has to be tilted toward rationalizations, otherwise, Føllesdal insists, any talk of psychological states loses its point. But within the theory, given the tilt, causal explanations are not subsidiary to, or derived from, or dependent for their meaningfulness on, rationalizing hypotheses in any sense other than that all elements of such a theory depend for their content on each other. This Quinean holistic interdependence does not prioritize any element over another, and so it is equally true that in Føllesdal's conception, while rationalizing interpretations must dominate the theory, they also depend for their content on the strictly causal explanations the theory invokes. The problem, however, is that the formulation of particular empirical generalizations of the latter sort

presupposes that we have a more or less firm, more or less independent grip on the concepts designating the kinds we thus link. But ideal interpretation is supposed to offer an account of what such a grip consists of, with respect to concepts describing thought and action.

If we imagine that we could step back from the characterization of IDA and ask what the items that interpretation reveals really are, then Føllesdal's humanitarian proposal may tempt us. For then we could imagine that both rationalizing accounts and empirical theorizing are providing us with indications, serving as evidence for the nature of the complex states we are trying to diagnose. But the naturalistic pretensions of the interpretive strategy are based on a refusal to allow a gap for ontology between vocabularies and their *denotata*. The interpretivist thinks that the only answer to the question of what content-states really are is an account of the vocabulary in which content-states are assigned. Once the question is allowed whether a vocabulary is *adequate* to the items it invokes, then the interpretivist loses this answer. The alternative is to regard the account of ideal interpretation as constitutive of the concepts applied, and hold that there is nothing more to be said about the relation between the nature of the members of the extensions of those concepts and the concepts themselves than what IDA tells us. If, however, we then go on to accept that IDA may invoke empirical, non-rationalizing generalizations in support of her theory-choice, we are giving up on our aspirations to offer, by way of IDA, a *basic* account. For now we abandon the idea that the vocabulary of action is *distinct* from the vocabulary (or vocabularies) of empirical law. And nobody could be misled into thinking that the interests embodied in a vocabulary of nomological generalization could be characterized by offering a methodology of ideal interpretation. In this case, the interpretivist strategy would not have succeeded in characterizing the vocabulary of agency and thought after all — it would characterize what I called above a pseudo-vocabulary. Once that is made apparent, the question of what thought and action might *really* be looms once more, to be answered, perhaps, in terms shaped by the interests that find expression in the pursuit of particular kinds of empirical theory.

To serve the pragmatist, the interpretive strategy must deliver a constitutive description of the concepts of action and thought. This means that we must not build into our account of the nature of these concepts and the interest they serve a reliance on generalizations that depend, as empirical generalizations do, on the availability in principle of a prior identification of the kind of states we are trying to characterize. If these considerations are sound, we have a conditional result: if the interpretivist strategy is to have a hope of meeting both its non-reductive aspirations as well as its naturalistic ones, it is going to have to be on the basis of RM. But of course RM may not be defensible. It may be that it simply cannot account for the grip on the

concepts of thought and action that we language-users constantly display. To show that it can, it is necessary to disarm examples geared to demonstrating that ideal interpretation by RM gives implausible results. What we need is not just an argument against some particular alleged counter-example. What we need is a general strategy directed at the root of anti-rationalist examples, which allows us to nip the humanitarian impulse even before counter-examples may bud from it. To provide such a strategy is the aim of the next section.

4. On Føllesdal on Urges and Attitudes: How Causal Generalizations Rationalize Persons

It will be helpful to separate two kinds of insight that typically are presumed to militate strongly against RM-based interpretivism and in favour of the humanitarian version. The first might be put as the claim that what can reasonably be imputed to persons must be informed by what we know about their modes of access to the world. Such knowledge may demand, for example, the ascription of erroneous beliefs to persons in a way that appears to be at odds with the requirements of RM. The second arises from the indisputable point that persons are in fact less than fully rational in their believing and desiring; we do not always believe or desire what we should believe or desire in light of other things we hold to be true and good. These I shall deal with in turn.

The first kind of consideration is emphasized by Grandy (1973).[28] Grandy points out that what we know about the way that the world impinges on us must necessarily constrain attributions to persons of views of how things are; "the causal theory of belief, " he claims, "accords much better with the principle of humanity than with the principle of charity."[29] What seems clearly right about Grandy's discussion of interpretation is the claim that "it is better to attribute to [a subject] an explicable falsehood than a mysterious truth." (1973, 445) Without impugning Grandy's treatment of his philosophical target, we may observe that this consideration does no damage to the view that RM governs the theorizing activity of the idealized interpreter. For one thing, even a rather vague injunction to maximize agreement or truth (by the light of the interpreter) applies holistically, and may in conjunction with the empirical evidence suffice to account for error attribution. Here the global scope of RM creates the necessary leeway.[30] A more revealing point, however, is that the main worry motivating Grandy's principle of humanity is in fact pre-empted in the very specification of RM-based ideal interpretation. Grandy is concerned to keep Quine's radical translator from attributing to subjects beliefs which, although true, it is highly implausible that they would have; beliefs which could not be a part of their perspective on things. The explicit point of ideal interpretation, however — as opposed, perhaps, to the more limited aim of the

constructor of translation manuals — is to provide a rationalizing description precisely of some field of causal relations on which supervenes the attitudes and the agency of some subject. It proceeds on the basis of a (revisable) view of what this field is, that is, of what the objects, events and relations are that fall under the scope of the theory. That view will involve among other things a theory of the nature of the subject's sensory connections to her environment. This causal theory, in turn, provides one constraint on IDA's comprehensive theory of perceptual salience, and it will generally (but not categorically) preclude assigning to the subject perceptual beliefs about objects with which she could not, by the theory, be enjoying any sensory connection. This means that IDA's choice of ascriptions is indeed constrained by causal theory. But there is no reason to think of this constraint as in any way qualifying or modifying the rationality-demand of RM. Rather, it contributes importantly to a *prima facie* delineation of the *scope* of the theory, by providing an initial fix on some of the items that should be accounted for, or that may be invoked, in the ascriptive theory delivered by IDA.

The second kind of consideration is the one I want to dwell on. It is nicely brought out in Føllesdal's discussion of the rationality-constraint. Føllesdal's point, on behalf of the humanitarian view, is that the right interpretation of persons in many cases will rely on empirical generalizations subsuming psychological states whose warrant is independent of their consequences for global rationality-assessments. RM-optimality cannot drive interpretation, since we can imagine cases where the strict application of RM gives a clearly less felicitous result than the application of a humanitarian principle. If this is the case, then we must have criteria other than rationality-maximization guiding IDA in selecting the theory that settles psychological and semantic ascription. *Ergo*, a weaker reading of the demand to minimize unintelligibility would be required than that delivered by RM. Admissible explanations of bad attitudes cannot be confined, as the rationalist would have it, simply to accounts that make them intelligible as misfirings of rational strategies and thought-patterns. While the humanitarian and the rationalist do agree that explicable error is preferable to inexplicable truths, the contrast between the views now emerges as a difference regarding what sort of an account of error to allow, in the context of interpretation, as explanation.

Føllesdal makes his point by means of an example he attributes to Patrick Suppes (Føllesdal 1982, 310). A young pupil with an attractive instructor very frequently comes up to the teacher after class, to ask questions concerning schoolwork. The sincere first-person account of this behaviour depicts it as a sustained attempt to obtain answers to questions regarding matters academic, an attempt motivated by a desire to learn. Undeniably, though, given what we know about persons in the throes of early puberty, it is easy to further specify the circumstances of the case in such a way as to make

the temptation to go beyond the first-person account positively irresistible. We may soon find ourselves explaining the behaviour not by invoking the student's professed desire to learn, but in terms of "urges adduced by the psychologist." (1982, 314)

Føllesdal elaborates this example to stress two points in particular. The first, positive, point is that "whenever [persons] experience [themselves] as carrying out [actions], what [they] do should be explained in conformity with the pattern of reason explanation."[31] (1982, 313) The second point is the one at issue; the case shows, Føllesdal believes, that we must reject the "normative methodological hypothesis" that in interpreting persons "we should always try to make [them] come out as rational as possible." (1982, 314)

Now, if we take this hypothesis to deny actual interpreters the right to support their efforts to figure out what is going on in the minds of their fellows by way of non-rationalizing causal generalizations, then I agree; we must reject it. If, by contrast, we take this hypothesis to be the claim that ideal interpretation is governed by RM, then we must insist on it. This distinction is essential. Certainly, as a principle alleged to be constitutive of a vocabulary, RM must be seen to be compatible with our actual interpretive practices. However, this does not immediately preclude endorsing humanity as a methodological ideal for *actual* interpreters — as a formulation of the strategies actual interpreters *ought* to follow. What we must show is that what actual interpreters thereby would achieve, is something IDA secures by virtue of RM. To show this is to give non-rationalizing causal generalizations of the kind Føllesdal urges good interpreters to invoke an integral place in a vocabulary the constitutive purpose of which is to reveal rational agency. Sticking by RM, the interpretivist must argue that the kind of causal psychological generalizations that we rely on to support actual interpretations can have that supportive role only in so far as they operate, in the ideal, in the service of rationalization. If we could show this, we would entitle ourselves to maintain what is the critical claim: psychological concepts are the concepts they are by virtue of their making evident in behaviour patterns that conform to the norms of reason.

Consider again the case of the pining pubescent. Føllesdal distinguishes various schematic possibilities regarding the relative explanatory power of the first-person account and the account couched in terms of urges. Even in the case where the first-person reasons offered "were neither sufficient nor necessary to explain [the] behaviour [and] the urges themselves were sufficient," (1982, 315) the reason-explanation must be part of our account of the event, otherwise we would not be talking about an action at all. But in this case a satisfactory explanation would also *have to* include the causal psychological factors, the teenage drives, those hormonally triggered urges. The implied case against RM here is that it would have us exclude the urge-

explanation, given that a rationalizing reason-explanation is readily available. But this would, *ex hypothesi*, be to opt for a bad explanation. Taking the pubescent to be acting on an acknowledged and reasonable desire may yield an account of what the student is doing that is, by common sense, inferior to one which has the student acting in self-deception, on the basis of unacknowledged urges.

I do not want to quibble about this common-sense verdict. Certainly there are cases where we quite reasonably take someone to be acting irrationally in the sense here at issue; that is, by taking them to be acting in blindness to their own actual motives. It is true, too, that such interpretations often will be supported by non-rationalizing generalizations. Let us stipulate that Føllesdal's pubescent provides us with just such a case. What, in the face of such concessions, can be said for RM?

Føllesdal implies that if we were to choose between the first-person account and the psychologist's story, we would be choosing between a rationalizing action-explanation and a non-rationalizing causal account of the behaviour. But there is another way to characterize the options. For what the psychologist does is to privilege another action-explanation, along the following lines; the pupil found the teacher very attractive, and, giving high priority to sex-related ends, designed a way to deepen and extend their personal contact. It would appear that viewing pubescents as weighing these kinds of ends heavily is a good way to make sense of much otherwise puzzling behaviour. And 'making sense' in this instance means 'making reasonable'. The effort to rationalize is still in view here.

Nevertheless, RM is in trouble, in so far as urges are invoked to justify this alternative action-description; the urge-explanation attributes a certain propensity to rank desires on the basis of physiological properties. It trades exactly on the kind of generalizations subsuming intentional states that appear to be thoroughly empirical in nature, and that therefore may yield ascriptions which conflict with the norms we have supposed, by RM, to constitutively govern the attributions of attitudes.

One way to dodge the problem is this. In our dealings with others, tracing in their behaviour the pattern of free, deliberate action is only one interest among many. There is a varying, but rarely insignificant, degree to which we care about people simply as objects. We should not assume that the features of concepts that allow agency to emerge are features of the terms by which we most efficiently describe and predict people for any purpose whatsoever. This raises the possibility that our ordinary dealings with people is conducted in a hybrid of vocabularies, where interests that may be at cross-purposes, or even directly antagonistic, find expression in different ways. Our common-sense intuitions, therefore, may be intuitions that express habitual ways of striking compromises between these interests. So it may be that when

we invoke common sense against RM we are smuggling in interests at odds with those that RM is expressly designed to capture — namely the interests we have that depend on our treating our fellows (and thus also ourselves) as autonomous and rational, i.e., as agents.

Whatever the plausibility may be of these considerations, they do not hold out much promise of a persuasive response on behalf of RM to the concerns that motivate the humanitarian position. This is because they straightforwardly deny an important humanitarian claim about the examples invoked; the examples show, it is alleged, that causal generalizations play an indispensable role in settling the *content* of the intentional states of the interpreted subjects. We could reply that the very notion of content thus being invoked is itself a hybrid notion, hostage to interests at odds with those connected with agency. This response, however, carries with it an air of stipulation so strong that we would be entitled to suspect that the subject is being changed. Certainly it would be an unattractive solution to anyone with pragmatist sensibilities; it is, after all, our ordinary attributions of intentional states, whatever work they do, that we are in the business of illuminating.

Granting, then, the point of the humanitarian examples, I want now to propose what I take to be a more fruitful line of response. Accommodated within the framework set by RM, cases such as urge-explanations will turn out have interesting implications for the nature of thought, implications that a naturalistic pragmatist should find independently attractive. Urge-explanation, as any explanation of a non-rationalizing kind that purports to ground attributions, may be seen as a way of coping with a fundamental tension in any interpretation of intentional systems; that between cohesion and scope. Roughly, the greater the number and variety of events that a single theory has to account for, the greater is the likelihood of anomalies and inconsistencies. When the strains are great, indeterminacy may increase up to a point where we begin to lose a clear sense of the contents being attributed. A possible response for IDA is to give up the unity of the theory — and with it, to some degree, the unity of the person — by allowing separate theories to account for different chunks of evidence. One imagines that these evidential chunks might overlap to a very large extent. The theories would differ, however, at least in what they discount as anomalous. They have, we might say, different focal points, around which meaning and belief are coherently rendered, focal points that will make them useful for different purposes. Each such theory would holistically deliver a compliment of attitudes and fix the interpretation of the agent's concepts, in accordance with RM. Such theories can also be brought together, in an *account* of a person, but when they are, the mode of their interaction would perforce have to be rendered in arational, causal terms.[32]

Nevertheless, we can easily imagine that the *construction* of such causal structures in the theorizing of IDA is governed by RM, that is, as constrained

by global rationality-requirements directed at the *person* being interpreted. There is, for example, the general demand that we bring all relevant considerations at our disposal to bear in our deliberations. Guided by such a norm, IDA would aim to minimize the number of consistent sets required to parcel out the ineliminable inconsistencies in any total set of action-descriptions, semantical interpretations and psychological attributions.[33] Here a demand for intra-theoretic consistency works in conjunction with a demand to keep at a minimum any non-rational inter-theoretic interaction in accounting for the sayings and doings of a single subject. Together, these demands — norms of reason — constrain the structure of the total *account* of the person arrived at by IDA.

What I have just now done, is introduce the idea of interpretational types. We actual interpreters rely on such types all the time, of course, often for ill in the form of pernicious stereotypes. My suggestion is, however, that even in ideal interpretation something akin to stereotypes is indispensable. Unlike us actual interpreters, IDA will have an ideally flexible range of such types, around which the various theories that make up the elements of the account of a person will be organized. For IDA, these types will be precipitated out as a result of the pressure imposed by the norms of RM on the behaviour of creatures like us.

Having introduced this notion of interpretational types, we must face the questions of what it is that is supposed to fix these types, to give them application to particular agents, and to allow IDA to keep track of their roles in her account of some agent. These questions point to the space that has now emerged for the kind of causal non-rationalizing generalizations that Føllesdal suggests must be brought to bear in interpretation. What such bodies of generalization do is designate an interpretational type, a particular perspective on the subject of interpretation, and anchor it to other identifiable features of agents. Take the pubescent-type, the type that we and the psychologist invoke when over-riding the first-person account of Føllesdal's pupil; to impose this type on a person is to structure a theory around a set of attitudes and intentionally characterized propensities, to give this set explanatory priority, and to *discount*, in that instance, evidence against it. In the case of the pining pubescent, where hormones are invoked, the type is defined in physiological terms. But this is not essential — generalizations linking intentional states of different kinds may similarly be called upon to support non-rationalizing attributions. In this case, too, as for example in psychoanalytic theory, we introduce interpretational types that cut across persons and are linked by causal generalizations to kinds of attitudes or dispositions.

What the psychologist does by treating the young student as an instance of a type imposing a certain value-hierarchy, IDA will do to us all. The point here is perfectly general, applicable not just to young piners in the throes of

hormonal urges they do not understand. To impose the type on some person, is to secure a set of attitudes at the heart of an interpretation, around which much, though not all, of the subject's behaviour can be usefully interpreted — in Suppes's imagined case as puberty-behaviour. What governs the interpretation here is still RM. What makes the behaviour the behaviour of a person transcending the type, are the causal generalizations invoked to identify the person as a person of that type. These generalizations will specify the attitudes and propensities that make up the core of the type, and link them to non-intentional properties, or intentional properties given by a theory not structured around that type. *Accounts* of persons now appear as networks of causally interrelated rationalizing *theories* of this kind, with each theory being geared to a type that gets projected onto the biological creature. Because each theory's particular focal point will be specified through generalizations that subsume intentional descriptions without also expressing the norms for their application, such ascriptive theories appear to be grounded in genuine empirical generalizations.

No one, however, should leave with the impression that such interpretational types represent empirically-discovered bodies of generalizations capable of delivering predictions of intentional states *independently* of RM. Take, for example, our reliance on hormonal states in imputing thoughts to pubescents. In the unlikely event that we were to develop ways of describing our pubescents in terms of easily diagnosable kinds yielding greater over-all RM-conformity than psycho-hormonal generalizations allow us to find in their behaviour, then psychological explanations in terms of hormonally induced urges would just die off. To that limited extent the issue between humanitarians and rationalists is perhaps an empirical one; the rationalist would predict that when there are systematic changes in the predicates that we invoke in causal generalizations subsuming psychological states, application of the new terms will generally yield greater RM-conformity than the ones being replaced.[34] Freudian theory is the paradigmatic example. The causal structures of the psychoanalytic soul rendered Fin-de-Siècle neurotics *more rational*, and thereby seriously modified the range of available responses to their neuroses; the liberating potential of psychoanalysis lies precisely in this fact. The rationalist would further hold that to show how a particular change in theory yields an increase in RM-conformity, would also be to explain that change. But that is not an empirical claim.

The role of psychological generalizations that are not expressions of the norms of reason, and therefore not constitutive of the concepts by which we ascribe thought and agency, is nevertheless to implement RM. By way of such generalizations, IDA both defines the interpretational types (cross- or sub-persons, if we like), and explicates their interrelations and conditions of application. In Føllesdal's example, we find that the imputation of a certain

desire-structure is justified by (indirect) reference to hormonal states. In other cases, we find behaviour-explanations invoking causal relations between differently focussed attributive theories — psychoanalytic models will provide specific examples. Explanation of action may invoke a particular theory with its defining focus, or it may invoke (causal) relations between interpretational types imposed on a single agent. In the latter case some degree or other of irrationality will be explicitly attributed to the agent. By building an account of a person consisting of distinct theories of sub- or cross-personal types, each with a specified warp, or a particular focal point, IDA accommodates the tensions in her evidential base by dividing it into more coherent subsystems. The central point is that the operative principle for IDA is still to make as much behaviour as possible as reasonable as possible. However, I claim, this principle may be upheld, when we are dealing with persons, not just by minimizing irrationality and error in a single theory accounting for all behavioural evidence. There is another axis of accommodation of the tensions that this evidence, as a body, inevitably produces; we may limit the scope of a theory, gerrymandering the evidence it is required to account for. We now treat the total body of evidence, the person as a whole, as subject to bundle of variously restricted, but causally related interpretative theories, each with a high degree of consistency and cohesion. We thus preserve the predictive and explanatory power of each theory, but we do so by sacrificing the psychological unity of persons.

Is this, however, really a sacrifice? Perhaps, in so far as the model of IDA's account of a person now emerging suggests an indeterminacy to meaning and belief over and beyond the ones familiar from Davidson's account of radical interpretation (1984). The added indeterminacy is significant; shifting between the theories that enter into IDA's account of an agent, we are not merely shifting from a concern with one action to another, or some region of an agent's psychology to another. We may find ourselves, as we move from one theory to another, individuating actions differently.[35] Not only may different theories account for the agent's behaviour differently in the sense that they may identify some action by different descriptions. Theories may differ even in what they characterize as an action. The clearest examples here are again probably psychoanalytic ones, where apparently or superficially viewed non-intentional behaviours may be redescribed in terms of the heretofore opaque intentions of a hidden or disguised locus of agency. But it can hardly be disputed that run-of-the mill folk-psychological practice also provides examples of context-determined shifts in the lines we draw between the intentional and the non-intentional behaviour of some person. The salient fact about this form of indeterminacy, however, is that it resolves into the context-bound nature of content. Unlike the indeterminacies of logical permutation or reference, or of the doxastic and semantic division of labour

within an attributive theory, the substantive indeterminacy of the identity of actions disappears as soon as we allow a context of interest to trigger a particular theory (or set of equivalent theories) amongst the possible theories embedded in IDA's account of a subject. Precisely because this strong form of indeterminacy leaves indeterminate something that makes a difference to practice — to how we respond to an agent, say — it is not a genuine indeterminacy. It only appears that way when regarded from a perspective that abstracts away from the varying contexts of interest and purpose that settle what we mean and think.

What is really at stake here, what my contextualist account of ideal interpretation puts under great strain, is the possibility of the reification of mental content. On the model of ideal interpretation I have proposed, the interpreter does not eliminate anomaly in behaviour. Rather the interpreter produces a set of devices, alternative theories, which allows us selectively to displace anomaly, deviance from norms of reason, and thus insulate behaviours or behaviour-patterns on which we may want for particular purposes to focus. The prevalence of conflict within the evidential base constituted by the actual behaviour of any entity of sufficient behavioural complexity to count as a person is universal. In the crucible of RM, such conflict forces upon IDA the strategy of interpreting differently circumscribed subdivisions of subjects, on pain of the dissipation of thought in a fog of indeterminacy. A consequence of this is that the patterns of reason traced by interpretation become multiply ambiguous. Reasonably determinate thought emerges only when an agent is interpreted as an agent *of some kind*, that is, *in some context, for some purpose*. Hence, ideal interpretation settles content only relative to contexts specified in terms of some subset of the various purposes, aims and interests we may have in approaching a subject as an agent. Such defining contexts may be just what come to expression in our characterization of the range of interpretational types by which we make sense of persons.

The aim of this section was to suggest how we can come to regard causal explanations as drawing their content from the application of RM, and how *prima facie* conflicts between RM and causal explanations disappear when we distinguish between normative principles for actual interpreters and the vocabulary-constitutive principles of ideal interpretation. The key move of the argument was the introduction of interpretational types as locus of the attributive theories of IDA. Causal generalizations specify the relations between various cross- and sub-personal types, and between such types and types specified in other terms, e.g., physiological ones. What remains fixed is that we explain and predict what persons do by rationalizing their behaviour, because it is only as rationalized that they act at all. What we have discovered, however, is that this very commitment dissolves the notion of mental content into a process of alternative and alternating rationalizing descriptions, each

representing some purpose-relative perspective on a person, a locus of agency. From the perspective of the pragmatist, campaigning for naturalization of our conception of persons by overcoming the metaphor of inner space and the reifications associated with the concept of mind, this should be a happy thought. Predicates designating mental states characterize aspects of agents in contexts of interaction with others in a shared world.

5. Bilgrami *versus* Nietzsche on the Contextuality of Content

It now appears that notions invoked in descriptions of ideal interpretation such as 'the perspective on the world,' 'the total theory of an agent,' or 'the totality of the behavioural evidence,' are misleading. They are misleading at least in so far as they suggest that there is a single, general perspective, defined by a general interest in agency as such, from which determinate thought-attributions and action-descriptions emerge. There is no such perspective, nor, hence, is there such a thing as *the* perspective on the world of IDA's subject. It is better to think of the idealization toward which these terms gesture as consisting in IDA's ability to form simultaneously an indefinite range of interpretational perspectives on some one subject, each of which may constitute its evidence differently. For IDA, with her account of an agent consisting of a set of causally related non-equivalent theories, there is no saying what the subject thinks or does *in general*; looking simultaneously through the various theories that go into an account of a subject, IDA would induce in herself an utterly blurring astigmatism. Determinate ascriptions of content and descriptions of action come only when the subject is regarded through one lens or another, that is to say, for some purpose or other. Thought and action emerge, as particular, interest-governed interpretative perspectives on behaviour are *actualized*. What our thoughts are and which actions we perform depend not only on what we do and what goes on in us and what the world is like, but quite literally on the particular perspectives from which we actually come to be regarded as engaging with the world. *A fortiori*, it depends on *there being* particular perspectives; I take the actual interaction of interpreters to be a condition of intentionality.[36]

The claims about content-attribution which underlie this conclusion bear some resemblance to one of the key commitments of the version of externalism that Bilgrami defends, the thesis of 'the locality of content.' I have emphasized the consequence of my view that the very identity of content-states becomes radically contextualized, dependent on particular explanatory situations and aims in a way that makes it interest relative and dependent on actual interpretation. Bilgrami does not accept this (1992, 238–241), but he has argued for a similar claim forcefully and at length. It will be worthwhile, therefore, to contrast my version of the thesis of the interest-dependence and

context-relativity of meaning with the locality-claim advanced in Bilgrami's appealing account of content.[37]

Bilgrami distinguishes what he calls the aggregative level from the local level of concepts. (1992, 10–13, 42ff) He suggests that meaning-theory, characterizing the former level, produces clauses that together specify, "all the beliefs an agent associates with each of his terms." (1992, 143) The concepts of such a theory, Bilgrami takes it, are useless for action-explanation. Explanation happens at the local level, and does not, at least not directly, employ concepts as specified by meaning-theory.

> Rather, what happens is that in citing beliefs (and desires) of agents at the level of the explanation of some piece of behaviour, we are distilling some beliefs out of the aggregate of beliefs summarized in any given clause or clauses at the meaning-theoretic level, and we are citing contents which are composed of concepts that are to be understood in terms of these selected beliefs. (1992, 143)

Bilgrami presents this view as a way to meet the claim that has shaped a significant part of Fodor's intellectual life, namely the charge that holists deprive content of explanatory power. Fodor's complaint (1987, 1994; Fodor and LePore, 1992) can be brought out thus. The explanatory point of a remark such as, "Because he wanted to get to Bombay, and believed that Bombay was the flight's destination" in response to the question, "why did Bilgrami board that plain?" trades on a regular connection between beliefs and desires of a certain kind and actions of a certain kind. For such regularities really to be regularities, we require identity of the predicates invoked across applications. But, since for Bilgrami (by his own admission — cf. 1992, 143) 'Bombay' is associated with what is probably a unique and unstable set of beliefs, the scope of the generalization shrinks to the point of vanishing and the explanation evaporates. If we endorse a holistic account of concepts, we get no explanation by attributing to Bilgrami 'Bombay'-thoughts; the concept of Bombay invoked here is useless, since, like Larkin's lading-list, it applied only to one man once. So holists, Fodor is convinced, have no right to feel satisfied by psychological explanation — and what is worse; if holism is true, nor does Fodor.

Bilgrami counters with the point that the requisite generality is preserved in local contexts, where only explanatorily relevant beliefs are invoked in the specification of the concepts that make up the contents attributed. To the worry that we now have two different notions of content going Bilgrami replies,

> There are two notions of concepts: aggregative and local. But there is only one notion of content because the aggregate level of concepts does

> not compose any contents at all. They are trumped-up posits, only there to acknowledge a larger pool of resources from which local concepts (which do go into contents) are selected ... .the only level at which any serious work is done by the notion of content is at the local level. (1992, 144)

My problem with this reply is not what it asserts. I find Bilgrami's notion of the locality of content congenial, and the non-reductive externalism in which it is embedded persuasive. The problem I have is that I am not clear on the motivation for the reply, as an answer to someone in Fodor's predicament; you need to say a lot more, it seems to me, or you need to say less, depending on who you are out to convince. I opt for less, and am not able to see why the following short reply to Fodor's worry will not do. Fodor thinks to himself, "When people want to go to Bombay and believe a certain flight will take them there, then, *ceteris paribus*, they get on the plane. Bilgrami desires and believes thus. That's why he got on the plane." Fodor feels satisfied. And why should he not? What really does the explanation here? Fodor's concepts; the generalization he relies on has application to all those to whom Fodor, by the evidence available to him, attributes the requisite belief-desire combination — *in his terms*, at that time. Period. If you are a holist and an anti-reductivist, and a naturalist to boot, this should do — you will be entirely unmoved by further worries about whether I (or Fodor) mean what I (or he) meant yesterday by those same terms, or whether, if we both agree on the explanation, we really aren't using terms differently. That may happen of course, but the only grounds we might have for thinking so is that the assumption of agreement leads to unexpected and puzzling uses of terms elsewhere; the hermeneutics of ascription is open-ended. But what the pragmatic naturalists will not allow any room for is a notion of meaning that is not fully exhausted by a rationalizing holistic characterization of speech and other behaviour. Such a behaviour-transcendent notion of meaning treats manifest behaviour as kind of indicator of underlying states. It opens the door for sceptical worries about identity of contents across agents and time-slices of agents of a sort that may remain pressing *in spite of* our successful prediction of speech and other action. It should be dismissed.

This will undoubtedly seem, from Fodor's point of view, like sheer, brick-headed point-missing; if you think *reduction* — or at least the clearly perceived possibility of reduction — is required for the legitimation of causal explanation, then the above method for reinflating the scope of explanatory generalizations will be entirely beside the point. So the short answer certainly will not make Fodor more favourably disposed toward holism. Bilgrami, however, is no reductivist. And what I do not see is why, if the short answer does not work, Bilgrami's does better. This is because I cannot, even with

Bilgrami's careful guidance, see my way clear to believing that we may individuate the items that are contextually selected and activated through the local concepts, which for Bilgrami compose explanatory content, in a manner that avoids complicity with the full compliment of attributions expressed by theories of meaning at the aggregative level. If we may not so individuate them, then Bilgrami's answer to Fodor is spoiled, collapsing into the dismissive one I just offered.

The version of the thesis of interest- and context-dependency of meaning that I propose is in any case not intended to provide an answer to Fodor's worries about holistic individuation of the *denotata* of psychological predicates (though the paper as a whole is intended to help those not yet firmly set in their metaphysical ways be less compelled by the kind of assumptions which yield Fodorian worries). My concern is with the nature and justification of the generalizations that may be invoked in ideal interpretation, and only indirectly with the problem of their scope. But my difficulties with the two-level account, offered as a response to Fodor, is rooted in an attachment expressing itself also in the difference between Bilgrami's locality-thesis and my contextualism. My doubts are rooted in what may be called whole-hog holism. Whole-hog holists think, as I still think, that any attribution of content and specification of a concept in ideal interpretation involves projections giving *maximally complete* descriptions of the evidence — under some specification. So in fixing concepts locally in terms of selected beliefs, we make implicit reference to some such maximally complete description. The claim I stand by differs from Bilgrami's; as a whole-hog holist, I am made a contextualist by the fact that I deny that ideal interpretation yields any single coherent theory at all of all the beliefs associated with each of an agent's terms, no matter how metaphysically light-weight we make the status of such a theory. The resulting contextualism provides me with an opportunity to accommodate the use of what looks like straightforward empirical generalizations within what is in essence a norm-driven enterprise. This contextualization yields refraction into multiple theories also at the level of the kind of transient, super-fine-grained concepts that are the deliverances of theories of meaning. Context, on the model I suggest, does not select doxastic items from meaning-theory, it selects entire theories of meaning and belief.

No maximally complete description of the evidence is a definitive account of all the evidence there is. Still, the notion of maximal completeness is necessary, it seems to me, to forestall a crippling indeterminacy. However, this constraint operates, as I have stressed, within the context of particular interpretive aims and interest. Theories of meaning may well be trumped-up postulates, abstractions of available resources, but even such theories, such characterizations, are selected among by IDA in settling some context of agency or other.

Since my thesis combines whole-hog holism with contextualism of content, it is in one way more radical than Bilgrami's. It is Nietzschean, in a quite distinct sense in which Bilgrami's is not; it denies that contextually determined meaning and belief refer back to any sort of coherently specifiable totality of ascriptions at all.[38] IDA's *account*, as I called it, is an abstraction; it specifies thought and meaning only in relation to particular contexts. Still, though this Nietzschean version of interpretivism is undoubtedly too ... well, Nietzschean (some, though not I, would say "anti-realist") for Bilgrami's tastes, the view I propose is not immediately incompatible with the two-level structure that Bilgrami posits. Furthermore, it should be emphasized that to contextualize meaning by depriving intentional systems of unity in the manner I suggest, is not to deprive *content* of unity in Bilgrami's sense (1992, 3–4, 15–16). The point of ascribing content to states of agents, lies in the ability such ascription secures for us to offer a distinctive kind of characterization of behaviour. The separation of contents ascribed in interpretation from psychologically explanatory contents is a disastrous move; unity of content in Bilgrami's sense is essential. My Nietzschean perspectivizing of content does not damage this unity, since context-relativity of meaning and thought pervades both semantic attribution and psychological explanation. What we think and mean explains what we do. What there is not — even as an abstract, rarefied posit — is a unified subjectivity of which what we think, mean and do are *manifestations*, and in terms of which our behaviour might be non-interest-relatively categorized and explained.

6. Pragmatizing Reason: Stich *versus* the Interpretivist Strategy

Though I have relied on Føllesdal and Bilgrami for contrastive force in staking out my position, the disagreements I have focussed on refer back to significant common ground. I want now to address objections to the interpretivist strategy as such, levelled from a necessarily more distant perspective. These objections put pressure on the naturalistic aspirations that inform my version of the interpretivist strategy. They will provide an opportunity to elaborate explicitly the notion of reason that RM invokes, and thus, I hope, to assuage the worries that Rorty has about the anti-naturalistic implications of the language I rely on.

Stich has been a vocal critic of interpretivism, culminating in (1990) where he invokes a range of empirical evidence for the systematic cognitive failures of human beings.[39] This, Stich claims, is incompatible with the interpretivist's commitment to the impeccable rationality of agents. Where Føllesdal relies on the common sense embedded in actual attributions, Stich relies on science to impugn RM. But Stich would have no truck with Føllesdal's humanitarian *ersatz*. He takes exception to the idea that any normative principle could be constitutive of psychological states. To see this,

we must distinguish two elements in his critique. The first is the claim, documented by the experiments of empirical researchers, that human beings evince systematic defects in their reasoning.[40] People just are not as rational as the interpretivist makes them out to be. So if the interpretivist is making an empirical claim, it is false. It should be clear by now that this point by itself does no damage to the interpretivist position, since it does not entail that people do not act irrationally. The second element, however, is for Stich the crucial one. What Stich finds objectionable is that the interpretivist is not making an empirical claim at all. If the rationality of human beings were the sort of thing that could be established by a priori philosophical argument, this "would make nonsense of the empirical exploration of reasoning and its foibles" (Stich 1990, 18). Furthermore, Stich claims, making the possibility of intentional-state attribution hinge on the rationality of the subject also threatens the one dimension of epistemology that he sees much point to. "It would ... turn the effort to articulate and defend a normative theory of cognition into an arcane and academic exercise of no particular practical importance." These two allegations I will address in this section and the following one.

I am about to argue that the interpretivist strategy does not have the first of these two alleged, and allegedly unwelcome, consequences. But what could be said of the second? Here I think the interpretivist should bite the bullet, welcome the imputation and follow Rorty in making a virtue of necessity. Furthermore, understanding the rationale for this response will help us bring out the sense in which the interpretivist strategy serves naturalism. Before pursuing this point, however, I return to the first allegation. The question is; does the interpretivist strategy run afoul of the sort of empirical research into human cognition that diagnoses, in folk-psychological terms, systematic propensities to err or sustain cognitive illusions?

Empirical evidence appears to show that in certain replicatable situations, faced with specific kinds of cognitive tasks, human subjects deviate in systematic ways from what appears plainly to be the desired result. Are interpretivists forced to take a line on this research that places them at odds with what scientists say, and what we folk should want to say, about these cases, and the manner in which the results are established? We may note, initially, that some disagreement persists among theoreticians of cognitive psychology about how to characterize the results of empirical research into human cognitive processing (see e.g. Manktelow and Over, 1993). So for example Gigerenzer (1993) argues that standard cases rest on interpretations of the tasks that are not mandatory; the appearance of cognitive illusions can be made to go away by paying attention to the interpretation that researchers implicitly are placing on, e.g., the laws of probability that underwrite their diagnosis.[41] Johnson-Laird and Byrne (1993), arguing the other side, spend several pages rehearsing what they take to be Stich's case against those who

doubt that empirical results can tell us how (ir)rational we are. One lesson a philosopher might be forgiven for drawing from the persistence of such argument is that amongst cognitive psychologists, as with Quinean radical translators, one theorist's evidence for irrationality is another theorist's evidence for the poverty of the former's model.

Let us, however, set this lesson aside, and not engage in what Johnson-Laird and Byrne call "heroic denial" (192) of the phenomena. It is difficult to feel anything but sympathy for their efforts to clear away the aprioristic or transcendental arguments of armchair thinkers to substantive conclusions about how good human beings are at cognitive processing. It is not, after all, as if there is no historical precedent for the kind of philosophical legislation that Johnson-Laird and Byrne impute to the "proponents of rationality." It is not good exegesis however, to lump interpretivists like Dennett and Davidson in with this kind of aprioristic, legislative philosophical practice. For, indeed, no question relevant to the viability of the interpretivist strategy is settled if we grant, as we surely should, that experiments do in fact reveal stable and persistent patterns of cognitive error. Nobody thinks people do not sometimes reason erroneously, and nobody ought seriously to entertain a theory of intentional states that requires us to deny that we can investigate the kinds of reasoning-mistakes that people are prone to making. The issue is what kinds of patterns it is we thus uncover, and how we ought to characterize them.[42] The key point is that the kinds of systematically occurring mistakes and persistent cognitive illusions that researchers report need not be regarded as indicative of irrationally maintained cognitive strategies, even if experimenters have isolated task-structures in relation to which those strategies lead us into error.

What is implicitly at stake here is the scope of the normative diagnosis, as well as the source of its normative force. A person may be convicted of irrationality only on the basis of a judgement of how his or her states of mind are related to one another (cf. Davidson 1985b; Stich 1990, chap. 2). A response to some cognitive task may be taken as indicative of irrationality only in conjunction with other states attributed (perhaps implicitly) to the subject. But then, if we are judging the rationality of the person's cognitive performance, considerations other than the validity of the particular piece of reasoning in question may be drawn into the picture. The *global* scope of IDA might well show it to be the case that these mistakes systematically occur, in suitably contrived circumstances, as the result of cognitive procedures that in fact are optimal, in a perfectly good sense. They may be procedures we should have deemed it irrational for creatures like us to abandon or modify (had we had a choice — which we often do not).[43]

Imagine that a cognitive procedure, one that in a certain type of case consistently brings subjects into a state of cognitive illusion, is triggered by a definite set of parameters. Though we can empirically locate kinds of cases

where the procedure and the set of parameters together cause a cognitive misfiring, this does not in itself show that we ought to modify either. The conjunction of judgements which expresses the cognitive illusions, however assailable they appear when regarded within a narrowly circumscribed context, do not in themselves support an inference to a global judgement that the subjects are not thinking as they ought to think. Errors and illusions certainly represent failures on the part of agents to live up to norms. However, these infractions must be viewed in the context of all the evidence available to IDA, evidence that, from the point of view given by the experimental context, would constitute collateral information. And it may well be that, in the light of such "collateral" information we actual interpreters would come to see the suspect judgements as results of a cognitive procedure we should deem it unwise to abandon. In that case, we may no longer be entirely comfortable with the claim that those mistaken judgements represent errors that the subject *categorically ought* to endeavour to put herself in a position to avoid. The global perspective of RM leaves IDA room to conclude that the norms in relation to which error is diagnosed may be trumped by metanorms, norms guiding a person's choice of cognitive strategy.

While experiments pinpointing the circumstances under which we tend to make certain kinds of error yield powerful constraints on proposed models for human cognition, it is not easy to see how such research can directly support firm judgements about the extent to which we, as creatures, live up to the ideal of rationality. The reason is that it turns out to be hard to agree on how to settle in advance the question of what kind of cognitive strategy the subject ought, in a given case, to employ. We could of course stipulate that a rational cognitive strategy relative to any given kind of task is one that consistently (and with maximum relative efficiency) yields what we agree to be the right result for that kind of task. (Setting aside, as we have been, the difficulty that arguments like Gigerenzer's may raise for particular conceptions of the right result.) By that standard we are, indisputably, less than fully rational. However, we should then also have to say that this is a kind of failure of rationality that it may be quite irrational for a natural creature to seek to eliminate. The diminishing marginal disutility of certain kinds of mistakes is such that the cost of developing or exercising cognitive strategies that would ensure their elimination is entirely out of proportion to the gain (if any) to be had from being thus error free.

The point I have just made is an instance of a familiar and perfectly general point about the rationality of real (and thus finite) creatures. But of course, those who claim the evidence shows that human beings systematically display irrationality will have an immediate retort. All this establishes, they will point out, is that human irrationality can be *explained*, something that certainly was not in dispute. That we as a kind are irrational in the particular

ways we seem to be may reveal something about the constraints under which nature has tweeked our cognitive strategies and capacities.

The point to notice, however, is that the explanation of the phenomena offered is one that *rationalizes* the apparent instances of irrationality in just the sense that RM requires; it renders them the natural outcome of the application of procedures to which *creatures like us* quite possibly *ought* to subscribe. (Or, as the case might be, by showing them to be the result of cognitive mechanisms we ought to be pleased that evolution has endowed us with.) The explanation does this by showing that the procedures or mechanisms are good for creatures like us. We might press the point by saying that how rational it is rational for us to be depends on the contingencies of our creature needs and interests and on the features of the environment within which we pursue their satisfaction. Looking for a more felicitous, less paradoxical way to put it, we should say that the idea of pure rationality, conceived as explicatable in terms of formal principles, is an idea that for the purposes of rationalizing interpretation of behaviour is without categorical normative force. When we as actual interpreters assess our cognitive practices for their rationality, any genuine critical force our judgements may have derives in the final instance from other aspects of our practices that we are at that moment, in the form of substantive normative principles, implicitly privileging as measures of worthiness. When IDA judges according to RM, what is required is an assessment of the contribution of particular aspects of practice or behaviour to over-all agent goodness — an ideal interpreter is, as Dennett (1981c) suggests, in this particular and substantial respect an ideal representative of ourselves.

Interpretivists are routinely chastised for refusing to come clean about what exactly it is they attribute to us all when they make rationality a condition of having mental states. This, we can now see, is because there is on their view nothing, a priori, to come clean about, except that to be rational is a very, very good and important thing to be. Interpretivists are staunchly anti-reductivist with regard to the notion of rationality that IDA implements; when we empirically investigate human cognitive capacities and strategies, we might discover all kinds of interesting tendencies and results — but there is no fixable, explicatable notion of rationality against which we can measure such findings and draw conclusions about the degree and distribution of rationality of human beings as a kind. Indeed, there is nothing in the pragmatist's interpretive strategy that suggests we could not come to adjust our assessments of rationality as a result of empirical study of our cognitive capacities.

We must reject the interpretation of the interpretive strategy that sees it as an a priori philosophical argument to a substantive conclusion about the quality or value of our cognitive procedures. What underwrites the connection between rationality and psychological attitudes is itself a species of naturalism; our conception of what we ought to be doing in the way of reasoning leads us,

as Dennett once put it, "eventually to a consideration of what we *in fact do*." (1981c, 98). It must be emphasized, however, that such considerations of our actual practices afford us no basis for a reductive account of rationality. Any gloss — or analysis — of 'rationality' represents some particular application of our cognitive practices to themselves. Whatever normative force such a particular application has, inevitably derives from attachments to aspects of our actual cognitive practices. These attachments, in turn, can be rooted nowhere but in experience, in the interaction of our creature need and interest with the environment in which we function. Perhaps one day it will be unnecessary to add that this does not mean that these practices cannot be meaningfully criticised or reformed — it implies only that they cannot be assessed wholesale, by some standard not of our own experiential devising.

Let us now turn to Stich's charge that the interpretivist's strategy of linking meaning and intentional states to the notion of rationality pre-empts the point of "a normative theory of cognition." Stich, a self-styled "normative pluralist" about reason, still wants, it appears, to be able to come up with a monistic evaluation schema for cognitive strategies, in terms of which, perhaps, the legitimacy of a plurality of incompatible strategies may be established, relative to the needs and interests and environments in which creatures do their cognizing. A burden, infamously, of Stich's argument is that 'truth' and 'rationality' will not play any explanatory role in this evaluation schema.[44] With respect to a representationalist conception of truth, Stich presses the question: "If *that* is indeed what it is for a belief to be true, do you really care whether your beliefs are true?" (1990, 22) He arrives at "a consistently negative answer," and concludes, "There is nothing special or important about having true beliefs." (1990, 24) At this point, however, the proponent of the interpretivist strategy might accept the conditional in Stich's rhetorical question but retain another option. In the spirit of pragmatic naturalism, the interpretivist should reply that we might surely stick with 'truth' and reject the representationalist psycho-semantic analysis. The result is that Stich's *modus ponens* becomes the pragmatist's *modus tollens*.[45] Because pragmatic naturalists will view 'true' and 'rational' as approbations like 'good' or 'beautiful', they will see them as flexible, fragmentation-proof sorts of notions — notions employed primarily to signal our approval of such cognitive and linguistic practices as we come to endorse in light of our changing and multifarious collection of particular aims and our evolving conception of what we want to be like. Here the interpretivist out-pragmatizes Stich's pragmatic theory of cognition. The question the pragmatist asks is, in effect: why should we feel compelled to monism at the meta-epistemic level? One answer is that it is a presupposition of the brand of epistemology to which Stich remains explicitly committed. But the pragmatist's reply to that, surely, is, "then so much for epistemology." Nor is this an entirely flippant answer. One might

offer it while indeed agreeing with a very great deal of the substance of Stich's diagnostic arguments against analytic and anti-sceptical epistemology, and also with the spirit of his normative cognitive pluralism. In particular, one may agree that there is no substantive content to the idea of "intrinsic epistemic virtue," (1990, 24) doubting along with Stich whether anything can be made of the thought that there are ways of reasoning that are inherently good irrespective of our contingent needs and interests. From here, however, a pragmatist should go on to doubt whether anything that is at all useful to the conduct of particular inquiries could be gleaned at the level of abstraction at which "a normative theory of cognition" must be pitched.[46] Stich does not seem to doubt this. Instead, having rejected, with due pathos, the idea that 'truth' and 'rationality' are concepts one may invoke to specify or explain what virtuous cognition amounts to, Stich then seeks an alternative conceptual framework within which to pursue such an explanation, precisely in the hope of resuscitating normative epistemology recast now as a quasi-empirical discipline (1990, 28). Embracing a full-blown, full-blooded naturalistic pragmatism, however, the interpretivist ought to reject this very project. The full-blown pragmatist will accept Stich's dim view of the prospects for normative epistemology based on a proper understanding of the content of 'truth' and 'rationality' but will not base this view on the claim that it sometimes is not good to be rational, or that truth, when properly analyzed, turns out to be something we should not obviously require of our beliefs. The pragmatist's scepticism toward this sort of epistemological project arises instead from the view that 'rationality' and 'truth' are not notions that have substantive normative content at all independently of our concrete evaluations of particular instances of cognitive virtue, evaluations which must remain, as Stich also insists, interest and purpose relative.

We Rortyans eagerly nod when Stich tells us that "appeals to rationality, justification, and the rest" cannot serve in any substantive sense "as final arbiters in our effort to choose among competing strategies of inquiry." (1990, 21) But we must dissent when Stich goes on to say that we are thereby "in effect, denying that rationality or justification have any intrinsic or ultimate value." The point, rather, is that *nothing* can serve as *final* arbiters in our effort to choose among competing cognitive strategies. Nevertheless, truth, rationality, and justification could no more fail to be intrinsically valuable than the good could fail to be intrinsically choice-worthy. What ensures this is also what guarantees that *analysis* of these concepts which aims to abstract away from malleable interest and contingent commitment, is normatively sterile — or self-deceptive.[47] And it is, importantly, what ensures that no question of empirical substance regarding the nature of our cognitive lives is begged by those who, pursuing the interpretivist strategy, treat RM-governed ideal interpretation as making explicit the nature of the concepts employed in folk-

psychological ascription.

Let us take stock. Interpretivists are not, contrary to Stich's first allegation, concerned to explain away the findings of cognitive psychology. How prone we are, as a kind, to making various sorts of cognitive mistakes is certainly an empirical question, as is the extent of our ability to learn to overcome such tendencies, or to compensate for them. So, too, are the extent of, and the causes of, variation in these regards amongst members of our kind. With respect to these empirical issues, the interpretivist construes rationality as a second-order category; particular kinds of error of reasoning do not *by themselves* indicate any particular degree of global irrationality. The global rationality-judgements of IDA express a view not only of the relation between psychological states and processes, but also of the relation between these and the constraints and needs and interests that provide the context in which these states are formed and in which such processes operate. Such global rationality-judgements are the ones on which IDA is instructed to rely when evaluating her candidate theories and accounts. And such judgements are simply not settled by the specific patterns of error that research psychologists reveal. There is, then, no conflict between the project of empirical investigation of particular cognitive mechanisms and the commitment of the interpretivist strategy to RM. Indeed, the interpretivist would claim, what gives us a firm grip on the patterns of error diagnosed by the psychologist, what gives us confidence in the identifications of the intentional states on which the formulation of any such diagnosis relies, is precisely their compatibility with RM as a globally regulative principle. The patterns of error traced by empirical cognitive psychology owe what sharpness they have to the possibility that just such errors may, from the global perspective of IDA, be good errors to make for creatures like us.[48] Nothing that empirical cognitive psychology could uncover would, unaided by metaphysical commitment, be capable of damaging this claim.

It emerged in Section IV that action-explanations may turn on non-rationalizing generalizations subsuming the kind of intentional states that we suppose to have caused the action. Neither from this, however, nor from the viability of empirical tracking of cognitive error-patterns, does it follow that we can empirically determine the extent to which we as a psychological kind are or fail to be globally rational in the sense required for the application of RM. The latter possibility is what the interpretivist must reject, as a possibility that is ruled out by the strong constraint expressed by RM. This rejection issues from the interpretivist's conception of the rationality-judgements on which we make ideal interpretation turn. Such judgements are, to condense the matter, expressions of a dynamic, evolving cognitive meta-practice of idealizing projection of what we actually find ourselves to be doing in the way of thinking and desiring.

Stich's second allegation was this. If we accept the interpretivist's conception of what it is to have psychological attitudes then we can no longer envisage an enterprise that aims to be normative with respect to our epistemic practices in general. This accusation, I have argued, is fair, and trades on a point the interpretivist should be pleased to concede. Conceiving of reason as the pragmatic naturalist does, any characterization of rationality or of warrant sufficiently abstract to appear philosophical will, by virtue of this fact, be normatively impotent. It will not tell us how to acquire fewer false beliefs, or desire better things, or act more wisely. To interpretivist ears, Stich's proclamation of the fragmentation of reason sounds like the final disillusion of a would-be essentialist, a reluctant lament for the fragmentation of epistemology. The latter is something that naturalistic pragmatists, particularly "vulgar" ones, should not feel inclined to grieve.[49]

7. Naturalism and Reduction

Stich, we have seen, thinks interpretivists are rushing in where all but scientists should fear to tread.[50] In this he is representative of a large group of philosophers of psychology. These are theorists who have taken to heart the Quinean view that the way to bring about the naturalization of some domain is to bring it under the scope of natural science. This commitment also comes to expression in Stich's view that interpretivism must be wrong, because it entails a manner of individuation of psychological and semantic states that renders them scientifically quite pointless.[51] The conceptual limits to irrationality that Stich believes fall out of the interpretivist strategy are "profoundly uninteresting" (Stich 1990, 51). "It is," Stich thinks,

> an observer-relative, situation-sensitive constraint that marks no natural or theoretically significant boundary ... Plainly, the demarcation between states that are intentionally describable and states that are not is going to be vague ... it will not be stable, or objective, or sharp. (1990, 51–52)

This distinction "is not one that divides nature at its joints."[52] The predicates characterized by the interpretivist strategy are thus useless to the cognitive psychologist. Stich might be willing to grant that the interpretivist strategy may plausibly be said to catch central features of the vocabulary of the folk-psychological attitudes; but then, we should conclude, so much the worse for folk psychology.

A similar concern is expressed by Fodor and LePore (1993b) when they argue, in intended *reductio*, that for all that the interpretivist could tell, it may be that we do not have beliefs at all, but simply *schmeliefs*. Now, schmeliefs

are,

> ...propositional attitudes exactly like beliefs in their functional roles, their qualitative contents (if any), and their satisfaction conditions, except that they are *not* analytically constrained by the principles of charity. To make matters worse, it might be supposed that it is *nomologically necessary* that schmeliefs are mostly true (mostly rational, or whatever) ... Then, *ceteris paribus*, the *only* difference between a creature's having beliefs and its having schmeliefs would be that, in the latter case, there are logically possible world in which what the creature has are mostly false, and in the former case there aren't. It might thus be really *quite* difficult to tell beliefs and schmeliefs apart. (1993b, 75)

Here Fodor and LePore gently mock the interpretivist for characterizing predicates that will not serve in a science of behaviour; for such a science it just could not matter whether a cognitive state is a belief or a schmelief. And while Fodor — with Granny looking over his shoulder — is unwilling to give the interpretivist the concept of belief (etc.), the point here is really the same as the one Stich relies on: making intentional states out to be intrinsically normative is to disqualify them as entities that could figure in any genuinely scientific account of human behaviour, since such an account does not need — indeed, could not be sensitive to — the distinction the interpretivist wants to draw between normatively constituted intentional states and other (putative) cognitive states.

These objections highlight the worry that interpretivism cannot satisfy the demands of naturalism. Since the interpretive strategy renders the vocabulary of thought and agency in terms irreducible to predicates that will allow a nomic account of human behaviour, it must be rejected. From this perspective, if you agree with the interpretivist that the strategy illuminates the concepts of folk-psychological practice, then this simply shows that folk-psychological states are not to be taken seriously (Stich). If, by contrast, it is your credo that these states are to be taken seriously, then, from the same perspective, it follows that the interpretivist must be simply wrong about the concepts of folk psychology (Fodor). In either case, you are taking it that the ontological fate of the reifications of folk psychology is separable from questions of what we as actual interpreters achieve by employing them and why we want to achieve those things — you are taking it that there is a substantive ontological fact here to be settled, one way or the other, by the success or failure of reductive proposals. On this perspective, the significance we ought to afford the vocabulary of agency — its "ontological status" — is a

function of our ability to link it up with a vocabulary of science. It could in principle be that in spite of its utility this vocabulary is actually ontologically inadequate. It could come to stand revealed, by philosophy, as invalid.[53]

For the pragmatic naturalist, the argument runs in the other direction; the irrelevance of the prospects of reduction to the run-of-the-mill purposes and interests served by our vocabulary of agency suggests that the naturalisation of this vocabulary has little to do with the supposed philosophical validity that reduction is alleged to provide. Consider the kind of dissatisfaction that Dennett's version of interpretivism often provokes. Reading Dennett, we quickly form the impression that to have beliefs and desires is to be predictable from the intentional stance. One might go on to think that this makes the vocabulary of beliefs and desires, in Dennett's words (referring to frog psychology), "a practically useful but theoretically gratuitous short-cut ... ." (1987c, 109). And then, impressed with Dennett's explicit disavowal of any principled philosophical distinction between frog-psychological states and human-psychological states, one might think that folk-psychology is simply a place-holder for a more enlightened, empirically adequate conceptualization.

Certainly Dennett has flirted with this view. And even when he explicitly retreats from it (1987b, 1991a) his critics often try to pin him to it. The intentional stance seems to Dennett's critics to make at once both too much and too little of the attitudes. The sheer, contingent fact of predictive success just seems too feeble a basis for a claim to realism of *any* sort; it is a fact that cries out for explanation, and it is here, among the terms of possible explanation, that the ontological action is. Such explanation might provide terms for a grounded realism toward the attitudes, or it might display the ontological shabbiness of the vocabulary of folk psychology. But Dennett's strategy claims for itself the right to endorse the attitudes while insulating them from the success or failure of this kind of explanatory descent. For Dennett, it is enough that folk-psychological explanation works, that it gets us what we want. For his critics, this is irresponsible; while folk psychology may be here to stay, as long as this is just because no better means of prediction actually happens to come along, this is not ontologically reassuring. The thought that if we *were* to develop better predictive schemes then that would spell the end of folk psychology, that thought seems just too irrealistic — instrumentalistic as the charge typically has it — to be the sort of thought we want to have about our beliefs and desires.[54] What makes Dennett's views so unsatisfactory to such readers is that he simply dismisses the thought that realists and eliminativists alike so clearly intuit: that the ontological status of the attitudes must depend on the fate of attempts to characterize them by means of the predicates of an account that actually *explains*, in other terms, the predictive success folk psychology appears to provide for its user-group.

It is easy, perhaps, in thus objecting to interpretivism in the guise of

Dennett's intentional stance, to lose sight of the fact that the fate of the vocabulary of intentional states is not, on the pragmatist's view, confined to the question of which predictive strategy is most reliable, or detailed, or elegant or precise or accurate. As much at issue is the question of what it is that is to be predicted. What we folk (psychologists) care about, typically, is not how people move various parts of their bodies, but what it is that we do by so moving them. And, again typically, whatever predicates we settle on in our descriptions of bodily movements, these are predicates agents can satisfy by moving their bodies in slightly, perhaps very, different ways. Such differences we generally want the predicates of our folk-psychological vocabulary to be insensitive to. What makes different movements instances of the same type of action, are the interests that give applicability to the predicates explicated by ideal interpretation.[55]

In all cases, some interest(s) will give point to our typology, and in all cases, "multiple realizability" of kinds of behaviour in physical movement would seem to prevail. There are no such things as brute psychological regularities because there are no such things as brute bits of behaviour. The point isn't merely that we only care to predict when we have some motive, or that some of the things people do matter more to us than others — though this is undoubtedly so. The point is that we *cannot* predict, indeed that there *is nothing to predict*, except in so far as we care about some things rather than others, in so far, that is, as we have predictive interests of *one kind or another*. Psychological explanation and prediction is, necessarily, of behaviour of this or that kind, and the kinds here refer us ineluctably back to need and interest.

Dennett (1991a) reveals his pragmatist stripes when he defends the integrity of folk psychology precisely by arguing the irreducibility of the types of this vocabulary to the predicates of some other vocabulary. Asserting the reality of the patterns we trace with intentional-state ascriptions, Dennett does not so much retreat from instrumentalism as take the edge of it by arguing that no other instrument will do for these purposes. He denies, by implication, that the predictive aims of folk psychology are specifiable in terms that transcend the vocabulary, and against which it could, as a strategy, come up short. Once we follow Rorty and bring the individuation of the very items of prediction under the scope of the vocabulary-constituting interests, instrumentalism ceases to be the thin end of the eliminativist wedge.

If the argument in Section IV has merit, the identification of actions is not only interest-dependent in a general way; the nature of these interests is such as to make the identity of intentional states (and thus actions) dependent on actual contexts of interaction. There is no fixing the elements of the subjective perspective of an agent on the world as such. To see an item as an agent, then, is not only to see the item as autonomous with respect to the categories of empirical law. It is also to see that item as possessing a nature

beyond what any determinate attribution of thought will make explicit; where agents are concerned there is, to paraphrase Heidegger, always more being than theory. I suggest that this is a constitutive feature of the vocabulary of agency — i.e., a part of what it is to consider some item as an agent. This is a way to articulate the Nietzschean element that I made explicit in Section V. To endorse it is to preclude the possibility that any vocabulary of empirical theory could ever do the job for which we rely on the ascription of intentional states. But if reductive legitimation of this vocabulary is ruled out, how can the vocabulary of agency realize a naturalistic purpose of any kind?

Reduction, says the pragmatist, is a meta-tool of science; a way of systematically extending the domain of some set of tools for handling the explanatory tasks that scientists confront. Naturalization, by contrast, is a goal of philosophy; it is the elimination of metaphysical gaps between the characteristic features by which we deal with agents and thinkers, on the one side, and the characteristic features by reference to which we empirically generalize over the causal relations between objects and events, on the other. It is only in the context of a certain metaphysics that the scientific tool becomes a philosophical one, an instrument of legislative ontology. This is the metaphysics of scientism. It treats the gap as a datum, and it takes natural science (or some sub-set of it) to be the philosophically fundamental account of what kinds of items we may, in a respectable voice, say that there are in the world. Identifying the natural with the science-side of the gap and the unnatural with the psychological side, scientistic philosophers like Fodor and Stich set out to either redeem or reject the latter in terms of the former. Given the assumptions, this is what naturalism demands.

The pragmatic naturalist, by contrast, treats the gap itself, that which transforms reduction into a philosophical project, as a symptom of dysfunction in our philosophical vocabulary. Pragmatic naturalism does not aim at conceptual reduction, but at a transformation of those conceptual structures we rely on to sustain our sense of a metaphysical gap between those items we catch in our vocabulary of thought and agency, and those items we describe in our vocabularies of causal regularities.[56] It is in the context of this metaphilosophical project that the interpretive strategy as wielded by Dennett and endorsed by Rorty emerges as a naturalizing one. It is not merely non-reductive, it is anti-reductionist; it seeks to free us from those philosophical perceptions that transform reductive enterprises into tests for ontological legitimacy.

8. Naturalized Philosophy

We may get a clearer sense of the philosophical context in which interpretivism functions by considering the following provocative remark of

Davidson's: "I can imagine a science concerned with people and purged of 'folk psychology', but I cannot think in what its interest would consist." (1987, 447) This stands in striking contrast to the sentiments of scientistic philosophers. Is Davidson suggesting that a cognitive science as conceived by the Churchlands, or by Stich — or, for that matter, by Dennett — is inherently without interest; that it could be of no value? This would be an absurd view to take, and thus an absurd attribution. The point of the remark is not that this would be an uninteresting science, but that such a science, however interesting, would not illuminate the philosophical issues that Davidson takes himself to be addressing; it answers to *different interests*. It would be wrong to think Davidson means merely that such a science would not be relevant to his particular concerns, however. His remark surely is intended normatively, expressing a conception of what philosophical concerns are, of what the interests are that philosophical reflection should be responsive to.

What conception might lie behind the thought that a science of behaviour "purged of 'folk psychology'" is philosophically irrelevant? It is a conception that ties philosophy to an interest in practice. The conception, however, is not simply a matter of being responsive to the demand that theory must be made relevant to our practical concerns, of resonating to Marx's final imperative in the *Theses on Feuerbach*. The relation is constitutive, not imperative. A part of the philosophical context that gives point to the interpretivist strategy is the claim that behaviour emerges as purposive behaviour only in the vocabulary of folk-psychology; it is only by the terms of this vocabulary that (some) events emerge as instances of motivated action. The constitutive point of the vocabulary is to show up agency. We have seen some ramifications of this. One, which I emphasized in Section III, is that the vocabulary will be structured around concepts that insulate the members of their extensions from nomic generalizations. A second is that the vocabulary yields determinate characterizations of agency only *as it unfolds*; no room is left for the idea of action as a manifestation of an underlying subjectivity (sections IV, V). A third ramification is this. For the pragmatist, as we have seen, attempts to reflect upon what there is are not distinct from reflection upon the nature of our vocabularies. Because we illuminate our vocabularies by giving explicit expression to the interests we take them to serve, philosophy itself, even at its most abstract, becomes wedded to the vocabulary of action. Any attempt to reflect upon the nature of things of some kind brings us to the question why we (should) *care about that kind of thing*, and this question will immediately throw us back into the vocabulary of agency.

This makes it evident why a science of human behaviour that gives up "the vocabulary of folk psychology" would be *philosophically* uninteresting. This should not, clearly, be taken to mean that there are not difficult questions philosophers may ask about what we do when we do science — science of

human behaviour and other topics — nor that individual sciences cannot pose their own peculiar philosophical questions, nor that philosophers may not contribute fruitfully to the reductive enterprises of science. But for anyone who conceives of philosophy as having an ineliminably practical interest — for anyone who thinks that our attempts as philosophers to reflect on what there is and how things are inexorably refer us back to a context which also involves questions of what we should value and what we should strive to become — to leave behind the vocabulary of agency is not finally to find a way to solve (or dissolve) philosophical questions about creatures with *psyche*. Rather, what we will then have found is a way to sever any tie between our topic and human *praxis*. For pragmatists, it is by their relation to human practice that philosophical questions take such content and point as they have.[57]

I have just proposed a view of philosophy that emphasizes the distinctiveness of the vocabulary of intentional states, of agency, and which ties philosophy as an enterprise to that distinctiveness. This may seem to place me at odds with Rorty (1997), who invokes the lack of sharp individuation-criteria for vocabularies to doubt that the "gulfs" between intentional psychology and physics and between (as his example has it) biology and physics, respectively, are "not equally wide, and of the same sort." (?) In both cases, we are just distinguishing vocabularies on the basis of distinct purposes. For Rorty, no vocabulary, or division of vocabularies, is philosophically special or privileged. There is an important truth to this, but I think its significance may be slanted by Rorty's fear of reason. The truth is that there is no other measure for critical evaluation of what we do or want than other things we do or want; there is no critique or justification that transcends the contingencies of need and interest, contingencies that give our vocabularies their shape. Recognizing this, however, does not force us to give up the idea that philosophy has a constitutive relation to the norms of reason. To insist on this relation, in the context of the interpretivist strategy, is just another way of stressing the point that philosophy is reflection on *praxis*.

The Reformist Rortyan claims, on behalf of philosophy, that when we invoke norms of reason we are drawing on interestingly distinctive explanatory resources. Rorty, however, balks at such apparently privilege-dispensing talk. He makes his point through a commentary on Davidson's "Mental Events" (1970), polemically aimed at McDowell (1994). Rorty rightly suggests that Davidson's notions of heteronomic and homonomic generalizations presume an independent ability to individuate vocabularies, and therefore cannot be relied on in an argument for the distinctiveness of vocabularies. Nor is this Davidson's strategy, though; I take him to be using the concepts to express the conclusions he draws from what he proposes are the distinctive features of our psychological vocabulary. Nevertheless, if the explanatory purposes for which we deploy biological concepts can be met only in so far as we build teleology

into our descriptions, then those concepts, as well, will constitute a vocabulary distinct from that of non-teleological science. And then, as Rorty says, the differences between the "gulfs" are "of the same sort." This is right, of course — at a certain level of abstraction. That is to say, the differences between the vocabularies are the same in so far as both bio-physical and psycho-physical generalizations come out heteronomic (as do psycho-biological ones). If this is so, then Rorty is right when he suggests that the particular contrast Davidson (1991a) wants to highlight between the concepts of the intentional vocabulary and those of other vocabularies is not illuminated by the distinction between heteronomic and homonomic generalizations; not illuminated, that is, by the distinction between vocabularies as such. But this is because *that* distinction is exactly blind to the *different sort* of differences that we may want to invoke in *justification* of proposed distinctions between vocabularies. It does not preclude that there may also be interesting differences between these inter-vocabularic relations. Davidson's intent, I suspect, is exactly that we should see an interesting difference between the sort of conceptual features that may distinguish the biological or the geological from each other or from the chemical or the physical, and the sort of conceptual features that make the psychological distinct from all of these. This possibility is certainly available even if we treat both kinds of differences as constitutive of vocabularies. This would explain why Davidson (1991a), as Rorty points out, revokes his earlier claim (1970) that the distinctiveness of intentional concepts arises from the indeterminacy of translation. It also means it would be wrong to conclude that Davidson could not be urging us to find an interesting difference between psychology and non-intentional sciences generally. The difference in question may be billed as meta-vocabularic: a difference amongst the sort of differences that we, using the 'vocabulary' vocabulary, can rely on to distinguish vocabularies. Once philosophers assume, as Rorty urges, and as I have tried to do, the 'vocabulary' vocabulary, this is precisely what *philosophical* accounts of things will seek to illuminate.

A reason, one might suppose, why Rorty appears less eager than Davidson to emphasize the distinctiveness and indispensability of the vocabulary of agency, is that he is a great deal less dismayed than Davidson about the prospect of leaving philosophy behind. This supposition, however, would be mistaken — or at least misleading. Rorty strives to naturalize our conception of philosophical reflection by thinking of it as an adaptive activity of natural creatures. We should, he urges, learn to think of ourselves in terms such that there no longer appears to be anything *conceptually* or *philosophically* mysterious about our being embodied thinkers, or agents in a world of causes. The interpretivist strategy naturalizes precisely in so far as it frees us from worries about the "ontological status" of the kinds that constitute the *denotata* of our various ways of describing things. While Rorty's

naturalism is in important ways Quinian, it is not Quine's naturalism, nor that of Quine's more scientistic descendants. By resisting the scientistic urge that informs both realism and eliminativism, the pragmatic naturalist insists that questions of what sort of predictive vocabulary to apply when, and to what — or whom, are questions that by their nature will not be contained within the scope of theoretical criteria of theory-choice. As questions of vocabulary choice, such questions resist methodological resolution. Neither mounting scientific knowledge nor the increasingly sophisticated theoretical superstructure of methodology raised upon it by philosophy of science will, all by itself, tell us under what aspects we should care about things.

Nor, however, and equally importantly, is this the kind of question that philosophy should be attempting to derive answers to by settling questions of ontology. Such decisions, which are normative and clearly decisions of fundamental importance, retain for pragmatists an ineliminably practical element — in the sense that they cannot be extricated, by abstraction, from what are essentially experimentally derived considerations of what we think we want to be like, what we want our practices to become. But what pragmatic naturalists with one hand take away from philosophy — the idea of ontology (whether as metaphysics or natural science) as a substantive enquiry into the legitimacy of vocabularies — they return with the other; we are left with a conception of philosophy as aiding our practical and ethical deliberations, our experimentations, by imaginatively providing alternatives to what begins to look like conceptual hang-ups and fixed ideas ('intuitions'), and depicting altered self-conceptions for us to try out. On this view, the job of a philosopher is to make vivid how our practices might change if we were to describe things — particularly human beings — in altered vocabularies, or if we extend particular vocabularies into new domains. This intellectual practice is not so much a pursuit of truth as it is a pursuit of alternative perspectives on the relevance to each other of various ways of making truth-claims. It is exemplified by the pragmatic naturalist's promotion of the interpretive strategy.

The interpretivist strategy undermines (as I have argued in sections III, IV and V) the reification of mental content and of subjecthood. At the same time, the strategy also frees the notion of reason from the transcendental aspirations in which it has been embedded (as I try to show in sections VI, and VII), and makes a notion of reason available for a pragmatized conception of philosophy. These two aspects of the naturalizing effects of the strategy are related. Both follow from a characterization of a vocabulary of reflection that aims to extricate our notion of agency and personhood from the dualistic, dichotomizing elements in the conception of subject and object that have come to be dominant in the modern stage of the narrative that Plato launched. These elements are what condition the opposition between reason and contingent

creaturely need, and they are what makes ‘ontology’ — the reductive reconnection of metaphysically ranked vocabularies — appear both as a domain of substantive enquiry and as a pressing task. Some of these elements are, to our detriment, still powerfully entrenched in our common vocabulary of the mental. They are no less active in the tough-minded resolve of contemporary physicalism than in the species-aggrandizing conceits of the early dualists of the modern era. Although they are still shaping conceptions of philosophical problems and of the tasks of philosophy, these elements are not presuppositions of philosophical reflection. In seeking to give them the slip, Rorty is engaging in the distinctively philosophical project of providing a reasoned view of better ways of being human.

NOTES

1. Though he would probably have resisted the radical contextualism I propose in sections IV and V, as I develop the interpretive strategy I take myself in the main to be following Davidson. I will provide very little in the way of explicit textual defence of the readings I impose on him, and on the works of Rorty and Dennett. Since the assimilation I imply is controversial, this may seem an odd omission. However, if the burden I assume in this paper can be sustained, it suggests that a Rortyan reading of Davidson and Dennett has the virtue of motivating aspects of their views that critics have found unsatisfactory. It is not my concern here to defend exegetical claims of a more categorical nature.

2. This diagnosis suggests that Rorty’s is a highly dynamic project, with an impetus for constant revision of its own terms built in right at the core. By the same token, it provides a methodological hypothesis for an account of the changes in Rorty’s views over the last 30 years. Both his early physicalism and a somewhat more lasting tendency to think of rationality in algorithmic terms (and therefore to be hastily dismissive of the notion) may be fall-out from attempts to naturalize philosophical reflection in terms that later came to be undermined by that very endeavour. (cf. Rorty 1994, 126.)

3. From here on, I will often use ‘naturalistic pragmatism’, or simply ‘pragmatism’ (and related forms), to refer to this view.

4. See Fodor (1987, chap. 3) and (1994, 5–7); and Fodor and LePore (1992, 1993a, 1993b). Compare Fodor’s remark: “If aboutness is real, it must be really something else.” (1987, 97) Dretske has a similar view of the options (1988, 1995).

5. Quine, of course, softens his stance in later treatments of the attitudes. See Quine (1991) for a view not very different from Davidson's on the ontological status of the attitudes.

6. See Kim (1993), particularly essays 14 and 17. For Davidson's defence of anomalous monism, see Davidson (1993a).

7. Reformist Rortyans follow (the later) Carnap and decline explicit argumentation about what it really is to consider something to be real. By contrast, Fodor's fundamental problem with interpretivism is exactly that our *explananda* are rendered insufficiently real, at least in a sense of "real" allegedly required by Granny's touch-stone intuition, tutored — as in Granny's grandson's case — by a clear-eyed appreciation of the ontological presuppositions of our idea of what it is for scientific explanation to succeed. For Kim, to take another example, non-reductive physicalists fail to perceive the incoherence of their conjunctive position because they do not see that, fundamentally, what there is, is properties; if states and events are individuated by their properties, then, for any physicalist, a non-physical property of some event must be the very property it is just because it supervenes on exactly those physical properties upon which it does in fact supervene. This allows us quite plausibly to conclude that a minimum commitment of physicalism is that non-physical predicates refer only in so far as they are nomologically related to physical predicates. Similarly, Frank Jackson (2000) takes the view that physicalism commits us to the view that assertions framed in non-physical terms would be made true by states of affairs characterizable in the terms of physical ones.

8. This is a central notion for Rorty, as Brandom (2000) stresses. I will have more to say about it in section III.

9. For Rorty's scepticism toward the idea of a truth-norm as deployed by Wright and by Haack respectively, see Rorty 1995a, 1995c.

10. Hegel, Nietzsche, Heidegger, and Gadamer make versions of this move the focal point of their thought, and this is what makes these thinkers such attractive tool-boxes for Rortyans. Allen (1993) demonstrates the power of this approach in his historical account of the role of the concept of truth in philosophy. McDowell (1994) and Brandom (1994) are recent works which fully integrate this semantic historicism into groundbreaking displays of constructive philosophy. Confronting issues that are still often posed and debated in ahistorical terms, as if the questions were posed *sub specie aeterni*, Brandom and McDowell both show up the intellectual poverty of ahistoricism.

11. This strategy has long been explicit in Rorty; cf. Rorty (1967).

12. Also of Davidson's. See Davidson 1989a, 1989b, 1991b. Rorty (1994) endorses the point.

13. On this reading, Dennett's conception of philosophical strategy and of how philosophical practice may contribute to our betterment is similar to Rorty's. The differences between these two philosophers are best explained with reference to their respective therapeutic aims. Rorty's attempts to affect vocabulary change are explicitly grounded in a commitment to democratic liberal politics. Dennett's stem from the over-riding goal of making it possible for us to adopt a self-conception consistent with the picture of human beings that is emerging from natural science without moral or spiritual impoverishment. Both philosophers, I believe, would regard these projects as related.

14. I hasten to grant the truth of the suspicion that my reading of Dennett may be anachronistic; as a matter of exegesis, it may be a distortion to see the formulation of the intentional stance, particularly as developed in 1978 and 1981a, as informed by the ontological tolerance which is present in Dennett's recent writings. Thus it may be that the pragmatic response to the charge of instrumentalism that I eventually offer in the final section is one that Dennett at that time neither would nor could have availed himself of. For my purposes in the present paper, I can afford to be agnostic on this point.

15. Davidson typically does not put the point exactly as I have done here. Still, I think what I have just said captures the role of Davidson's notion of radical interpretation, *pace* Fodor and LePore (1992, 1993a). Fodor and LePore attack Davidson for *assuming* that the methodology of radical interpretation will work. Searching through Davidson's writings, they find proffered therein no reason to believe that radical interpretation is possible. Davidson's reply (1993b, 77–84; 1995) is to the effect that they are reporting on the results of a wild goose chase, one they are misled into embarking upon because they seriously misconstrue the dialectic of the interpretivist's strategy. In particular, they badly misdiagnose the relation between ideal interpretation and actual interpretation, as I have just set it out. What an interpretivist is committed to demonstrating, is that the explicit methodology of ideal interpretation (whether Davidson's radical interpretation or some other idealization) tends to end up with just the state- and meaning-attributions we ourselves think are appropriate in given circumstances. Put another way; we may innocuously assume that radical interpretation is possible, as long as we *do not* simply assume that radical interpretation is right interpretation.

Interpretivists may frame this issue in different ways, depending on context and purpose. Coming at it, as it were, from the left, one may determine that the specified methodology of ideal interpretation will yield some interpretation or other, and then set out to investigate whether the interpretation thus arrived at is one the folk would approve of. Approaching, alternatively, from the right, we may — less transparently, perhaps — assume that the outcome of ideal interpretation must be just what we would want it to be, and then go on to wonder whether thus successful interpretation, as governed by principles and constraints specified by the theory, is in fact possible. Perhaps it is fair to say that Davidson has at different times come at the issue from either direction. What he has not done — what an interpretivist cannot do — is *assume both* that ideal interpretation (as per specified constraints and principles) is possible, *and* that ideal interpretation yields the right interpretation. Someone who made both these assumptions simultaneously would have no need for arguments.

16. Davidson follows Quine and characterizes this last idealization — the ignorance-condition, as we may call it — with the adjective "radical." I think it is useful to emphasize also other dimensions of idealization involved in the construct which embodies the methodology at the core of the interpretivist's position. Hence my relabelling of what is essentially Davidson's construct.

17. By changing the crucial meta-attitude to be identified by the radical interpreter from that of holding-true to preferring-true, Davidson (1990a) explicitly unifies decision-theory and truth-theory in a single interpretational enterprise. Since attitudes of all sorts can be nailed down in terms of truth-preferences between pairs of indicative sentences, the formal constraints we impose by modelling a theory of meaning on a Tarskian truth-theory now force structure not just on the semantic theory, but on the ideal interpreter's account of everything she ends up taking as actions. A Tarskian truth-theory gives shape not just to the interpretation of language, but to interpretation more generally; it makes possible both interpretations of words and individuation and attribution of attitudes. See also Davidson 1980, 1995.

18. The claim so far is only that without some general constraints on the pattern of preferences that IDA observes, observation would be useless for purposes of theory-construction. I shall arrive at a somewhat less bland claim in Section VII; in the absence of such *desiderata* IDA would not only not be able to construe observed behaviour as evidence for a theory, she would not be observing *behaviour*, of any sort, at all.

19. This way of characterizing what IDA reveals is intended to

emphasize the gratuitousness of the idea that the interpretivist strategy leaves out of account, or may place one at odds with, a *bona fide* first-person perspective. For a vivid exposure of the superstitions that give rise to the gratuitous idea, see Bilgrami (1992, particularly 49ff and 225ff). "It is true," however, as Bilgrami notes, "that an externalism may be insensitive to the right constraints, in which case externalism would indeed be guilty of the charge of failing to capture the agent's point of view and, therefore, failing to get right what the agent really believes." (1992, 237) The right kind of sensitivity is ensured by the demand that all concept-fixing and content-assignment is governed holistically by the totality of such assignments given by some interpretation.

20. Speaking more strictly, we should refer to the set of optimal theories, to allow for different but empirically equivalent ways of describing the evidence. Complications are introduced in Section IV.

21. Bilgrami offers the following constraint as the corner-stone of his brand of externalism: "(C): When fixing an externally determined concept of an agent, one must do so by looking into indexically formulated utterances of the agent which express indexical contents containing that concept and then *picking that external determinant for the concept which is in consonance with other contents that have been fixed for the agent.*" (1992, 5) Constraint (C) emphasizes the hermeneutic dependence of any concept-fixing clause offered by the interpreter on the attribution of contents to the subject of interpretation. This is the heart of the difference between Bilgrami's view of content and those associated with Kripke and Burge, his principal targets. It forms the basis for Bilgrami's trenchant critique of "orthodox" externalisms, views which are not species of the interpretivist strategy. This polemical orientation is one reason why the relation between Bilgrami's proposal and the issue I frame in terms of the contrast between humanitarians and rationalists (Section III) is not transparent. Bilgrami takes (C) as an alternative to the humanitarian proposal that we seek to minimize inexplicable error. Certainly, Bilgrami's Constraint (C) differs from Grandy's proposal at least in so far as it guides the interpreter's selection between alternative *descriptions* of the objects to which some (indexical) concepts of the interpreted agent is being linked, even in cases where no contemplated alternative would result in the imputation of a false belief. (Bilgrami 1992, 7–8, 237) Bilgrami is surely right to insist that the role of the interpretation-constraint cannot be limited to guiding or restricting the attribution of error. Nevertheless, it seems to me that it is possible to raise the issues I will pursue (sections III and IV) within the context set by (C). Bilgrami, following the line of thought of Quine ("occasion sentences") and Davidson ("the simplest and most basic cases"), takes utterances containing

indexicals as entry points for interpretation. He also insists against orthodox causal externalism that even indexically-based assignments are holistically constrained; "there are no unmediated causal links" between environment and contents (9), no content-establishing causal links unmediated by other content-attributions. Precisely this important point, however, requires us to say something about the nature of the pattern that the interpreter is to make explicit beyond its internal coherence. The notion of 'consonance' invoked in (C) is to be understood in terms of formal and material inference relations; the constraint is free of empirical psychology. Constraint (C) guarantees the coherence of the pattern of inference relations constituted by the beliefs composed of the concepts attributed to an agent; it thus emphasizes the demand that we find the subject's means for describing the world coherent with her view of the world. RM also demands this. But I do not see that (C) by itself tells us — as RM does, and as principles invoking agreement, truth, or explicable error have all been intended to do — how to stabilize attributions of attitude-contents on the basis of the available evidence firmly enough to allow a reasonably determinate fixing of holistically determined concepts. In the dialectic I set up (Section III), the question of how to achieve that stability is just what distinguishes the rationalist and the humanitarian positions.

22. This is a worry Rorty has expressed in conversation; it is exactly the worry that I, on behalf of Reformist Rortyans, should like to alleviate.

23. Charitable proposals typically suggest maximizing agreement or truth. (See Føllesdal (1975) for incisive criticism of this idea.) Humanitarian counterproposals typically insist on modifications which are alleged to bring the attributive theory in line with what, intuitively, is the perspective of the agent under interpretation.

24. Cf. Davidson 1982, 1986a, 1989c, 1991a, 1992, for the evolution of this view. Jennings, in the preface to his remarkable account of disjunction, expresses a commitment to pragmatic naturalism in the philosophy of language thus: "If an instructively oversimple slogan were to emerge from my efforts and be offered as amicular advice to discourse generation researchers, along the lines of the earlier 'Don't ask for the meaning; ask for the use', it would be 'Mainly we emit sounds'." (1994, ix)

25. Even commentators with great sympathy for Davidson's views (e.g. Farrell 1994) think Rorty's retreat from ontology is a retreat from the constraints of the world. I hope my reading makes evident how wrongheaded this is. Rorty, following Davidson, takes thought to be a natural capacity of some worldly creatures. It is only in a world filled with the kinds of things we

generally think and talk about that thinking and talking could emerge as natural coping strategies.

26. The concept is ubiquitous in Rorty's writings (see Brandom 2000b). I can think of only two places, however, where Rorty considers the individuation of vocabularies (1989, 7fn; 1998a). Rorty (1989, 7fn) and Brandom (2000) regard a vocabulary as something that is suitable for translation. In what follows, I diverge. Certainly there is a sense of 'vocabulary' which fits this characterization, for example when we talk contrastingly of the vocabularies of Aristotle, Newton and Einstein. But I think that even in these cases, the sense of 'translation' is derivative. When we, as I do in this paper, speak of the vocabulary of mind, of the attitudes, of psychology, biology or physics, or when we speak of the vocabulary of norms, or of virtue, or of rights, or of New Age Spiritualism or Mahayana Buddhism, it sounds to me odd to say that we are thereby picking out suitable objects of *translation*. Certainly, we can translate things uttered in such vocabularies, but that is because they are utterances, and so bits of language, and hence translatable. Indeed, I don't find it the least bit odd to think that successful translation may be a sign of a shared vocabulary — just imagine a foreign New-Ager equipped with a pocket dictionary touring hip desert towns in Arizona. Conversely, we can be brought to realize that we are encountering some vocabulary we do not know when translation bogs down unexpectedly. What we then require is only in a derivative sense translation. Think of the bilingual professor translating from her copy of *Kritik der Reinen Vernunft* being asked by the eager but Kantless student to translate what she just said into "ordinary" English. What is really being asked for, what we really need in such cases, is an interpretation of the relevant practice, one which explains the norms governing the use of the terms by making clear to us what the constitutive commitments are — and thus telling us what the practice *is*.

27. To say this, it is not necessary for me to deny that we can sometimes *characterize* those interests in another vocabulary. But such characterizations are parasitical, in the sense that we rely on the vocabulary we are thereby evaluating to *identify* the interests we characterize.

28. The same point is behind Føllesdal's claim that "in a satisfactory theory of meaning there seems to be no way of avoiding the study of sensory experience." (1975, 40) Still, for Føllesdal, the connection between an account of sensory stimulation and a theory of meaning is certainly not direct: "Even though the impingements on my sensory surfaces may remain the same, what I experience may come to differ, as my beliefs and theories concerning the world change" (1982b, 559). Davidson (1990b) goes further when he doubts,

in criticism of Quine, that a theory of sensory stimulation would have any role at all to play in the project of IDA.

29. Grandy's principle of humanity says that "we have, as a pragmatic constraint on translation, the condition that the imputed pattern of relations among beliefs, desires, and the world be as similar to our own as possible." (1973, 443)

30. Several people have made this point, e.g. Evnine (1991), Malpas (1992).

31. The principle captured in the first point is precisely not intended to exclude further or complimentary psychological explanations involving non-rationalizing descriptions of the events. I read it not as a proscription at all, but simply as another embodiment of Føllesdal's commitment to the thesis that all action-explanation rests constitutively on rationalizing interpretation. For the principle is, I think, as near-analytic as the present context allows: since the state "experiencing oneself as performing an action" takes its content, like all other mental states, from its place in the hermeneutic circle of third-person interpretation, it is hard to see how we could fail to conform. Once IDA has attributed a description of an event under which the subject regards it as an action, she will have applied the pattern of reason explanation. The relation (within the theory) between such a description and the attribution of the relevant self-understanding to the subject is intimate indeed.

32. The model here is Davidson's treatment of mental division (cf. Davidson 1974, 1985a, 1985b). Cavell (1993) provides an interesting explication of Davidson on irrationality, and puts the Davidsonian notion of mental division to creative use. Without blaming her for the particular proposals I make here, I want to acknowledge a significant debt to Cavell's account of psycho-analytic concepts in Davidsonian terms.

33. Ray Jennings has suggested that paralogical notions provide formalizations of just the kinds of constraints I have in mind here.

34. Ideally speaking, that is. But the *ceteris paribus* clauses hedging this prediction are made rich indeed by socio-political interference.

35. This is the kind of indeterminacy that Dennett (1991a) stresses. Unlike Bilgrami, I think semantics and intentional psychology is rife with this kind of individuative indeterminacy.

36. As does Davidson, though he may not have been attracted to my way of arriving at the view.

37. Particularly given Rorty's view, quoted on the dustcover of Bilgrami's book, that, "Akeel Bilgrami has taken a giant further step along the path broken by Quine and lengthened by Davidson — the path to a radically naturalistic theory of meaning."

38. "No, facts," Nietzsche notoriously proclaims, "are precisely what there is not, only interpretations." (1968a, 481) With this and other formulations of his perspectivist view of truth (e.g. 1968b, III (12)), I take Nietzsche not to be proclaiming a wild-eyed scepticism or nihilistic relativism, but to be registering an anti-representationalist's complaint against the very idea of ontology. Much the same could be said about my perspectivist view of agency.

39. See also Fodor and LePore (1993a), and Dennett's reply (1993, 215ff).

40. See Stich (1990, 4–9) for a quick overview of some of the results of this research.

41. Gigerenzer, referring to the research of Tversky and Kahneman, discusses our tendency to commit what is known as the "conjunction fallacy." Provided with information about the hypothetical but now nevertheless quite famous Linda, subjects go on to rank for their probability two possible descriptions of Linda. One alternative is a conjunction (Linda is a bank teller and active in the feminist movement), and the other is one constituent of that conjunction (Linda is a bank teller). It turns out — sadly, one might think — that most of us are inclined to assign a greater probability to the conjunction. Worse, it seems that most of us, while happy enough to acknowledge that a conjunction cannot be more probable than one of its conjuncts, find it very hard to counter-act our tendency to reason in this way. Having the error pointed out to us in a particular case, does not seem to help us much in our next encounter with a case of this kind. (Gigerenzer 1993, 284)

42. Dennett makes a version of this point in response to Stich (1981a) and elsewhere. (Dennett 1981a, 1981b) Dennett's view is that to talk sharply about such phenomena, we need to retreat from the intentional stance to the design stance. On my model, RM gives content also to our diagnoses of irrationalities and cognitive error.

43. Stich seems to me exactly right about this — that rationality judgements are comparative judgements, and presuppose a context of specific goals, purposes and empirical limitations.

44. "Notorious" may be the better word; see e.g. Haack (1995). In what one with a polite euphemism might term a "spirited" paper, Haack lumps together Stich and Rorty under the catchy epithet "vulgar pragmatism."

45. Rorty (1992, 1993b) and Davidson (1991) are explicit here, invoking Putnam's naturalist-fallacy argument against the very project of providing an analysis of truth. Dennett (1981c) makes exactly this point when he refuses, in reply to Stich (1981), to commit himself to any particular analysis of 'rationality' on the grounds that it is "a pre-theoretical notion": "I want," says Dennett, "to use 'rational' as a general-purpose term of cognitive approval — which requires maintaining only conditional and revisable allegiances between rationality, so considered, and the proposed (or even universally acclaimed) methods of getting ahead, cognitively, in the world." (Dennett 1981c, 97)

46. Cf. Williams (1991) for an elaborate argument against the coherence of the project of doubting or legitimizing "knowledge of the external world."

47. Though this is not to suggest that such analysis may not have value, and indirect normative implications, in particular contexts, where particular ends and interests are at stake.

48. I have not relied on this point in my response to Stich since it assumes the point Stich places at issue, namely how we should conceive of the identity conditions of psychological states.

49. See Haack (1995) for a set of criticisms levelled at Stich from the perspective of someone who wants to defend the idea of epistemology.

50. Davidson anticipates this reaction to his brand of non-reductive naturalism: "Do we," he asks, "by declaring that there are no (strict) psychophysical laws, poach on the empirical preserves of science — a form of *hubris* against which philosophers are often warned?" (Davidson 1970, 216)

51. Here is Stich (1992): "The literature strongly suggests that those who want a naturalistic account of mental representation want something like a definition — a set of necessary and sufficient conditions — couched in terms that are unproblematically acceptable in the physical or biological sciences."

(1992, 260) Stich disparages this reductivist impulse, in a recognizably pragmatic spirit, distancing himself from projects like those of Fodor and Dretske. His pragmatism remains Quinean, though, in a sense that is shared neither by Davidson nor Rorty. This is because Stich's sceptical attitude toward the interpretivist strategy depends on treating science as settling what we are entitled to say there is. For the Rortyan pragmatist, there is *no* particular vocabulary, not science, not metaphysics, which has a special, legitimating role by virtue of a capacity to settle what in general there is.

52. Stich attempts to make vivid our appreciation of this point with a thought-experiment; a sequence of brains each (but the first, namely Stich's) of which is a computational duplicate of the former, except for one of the sentences in the "belief box." Considering the owners of the brains, Stich thinks that "when we attempt to describe these people in intentional terms (in a given context), we will be forced to divide them up into two radically different groups. The ones relatively close to me have intentionally characterizable states, the ones very far away do not. If the computational paradigm in psychology is on the right track, then this distinction, mandated by the chauvinistic principle of humanity, is without any psychological significance." (Stich 1990, 53) It is probably best to take Stich here to be speaking in the spirit of explication, and not argument. Still, it does seem remarkable that the passage invites us to assume that the syntactically characterized objects at a functionally defined brain-space settle which (if any) folk-psychological intentional states the creature whose organ the brain is might be in. With regard to Stich's sequence, what the rationality-maxim in fact would mandate, were we tempted to follow Fodor and insist on identifying folk-states with computational states, is that we refrain from identifying types of the former with types of the latter *across* brains — who is interpretable by whom would then just not be something Stich's case tells us anything about. But, of course, the interpretivist has no incentive at all to follow Fodor this way. On the contrary, as far as the interpretivist is concerned, her best buddy may well have a "belief box" that differs radically from her own, with entirely different syntactic objects in it. So also with "the principles that govern how these inscriptions interact with one another" (1990, 53); for the interpretivist, questions about how, and on what, your soul-mate's brain performs its computations simply are not at issue when you marvel at the astounding frequency with which you find yourself articulating each other's thoughts. Attitudes are fixed by an ideal interpreter which attributes them to persons depending on when they do whatever they can be observed to do (including what noises they make); undoubtedly, people's brains play an essential part in this. But it does not follow from this that the *brain* and its states are the proper subjects of belief attributions. In the context of a polemic against the

interpretivist, Stich's invitation is completely tendentious. That he issues it, is an expression of the hold of the assumption he shares with Fodor and countless others: the naturalization of folk-psychological states requires us to find ways of characterizing those states in terms of predicates licensed by scientific theory.

53. The kind of worry is expressed by Fodor (1987) and Dretske (1988). Cf. Stich (1991).

54. See Dennett (1987b) for some regrets about "instrumentalism" and other labels, a lament he carries further in (1993).

55. Similarly with empirical psychology: perhaps, as Dennett suggests, the job of psychology is to formulate regularities that serve in explanation of "the reliability with which 'intelligent' organisms can cope with their environments and thus prolong their lives." (1981b, 64) Then this explanatory interest would be what ultimately settles what counts, for some behaviour, as being an instance of that behaviour. Perhaps there are other ways of characterizing the explanatory interest of cognitive psychology that are no less plausible. Such a different explanatory interest might be sensitive to slightly different ranges of differences and similarities, and so classify behaviours differently.

56. By this characterization, McDowell (1994) is a pragmatic naturalist. I do not think he would object to this; indeed, in this paragraph I take myself to be following McDowell's lead. It is a central lesson of *Mind and World* (see particularly lecture IV) that if we are to "reconcile reason and nature" (1994, 86), we must exactly challenge those ways of thinking that make it appear as if reconciliation must take the form of reduction. The differences between McDowell's metaphilosophical stance and Rorty's are smaller than McDowell's appropriation of Kant might suggest. McDowell takes a much more optimistic view than does Rorty about how much of the vocabulary of modern philosophy can (and should) be successfully reformed through a naturalistic transformation of the vocabulary of mind; their therapeutic aims, however, are shared.

57. This point is perhaps easier to read into Davidson's writings than into Dennett's, but I should not think Dennett would be in serious disagreement with what I have just said about the point of folk psychology. Dennett remarks on the inescapable nature of the intentional stance in (1981a, 27). In (1991a) he also makes clear that he regards the individuation of the kinds of behaviour predicted from the intentional stance as interest relative. I

should, however, like to extend this point also to non-intentionally characterized behaviour — indeed to predicates that serve predictive regularities in general. Since I don't see what compels us to treat the idea of a basic law as anything but an abstract idealization, I do not see why we should believe that there must be a level of regularities where all questions of implementation become otiose — where, that is to say, all that could possibly be offered by way of explanation of apparent constant conjunctions is to say, with Fodor, that "God made it that way." (Fodor 1991) Indeed, Fodor's quip is highly apposite, since the idea of a finished physics and the idea of omniscience — the one (set of) coherent account(s) that accounts for all there is to be accounted for — are mutually supportive notions; doubting the point of one would seem to leave the other in serious trouble.

REFERENCES

Allen, Barry. *Truth in Philosophy* (Cambridge, Mass., Harvard University Press, 1993).

Barrett, Roger, and Roger Gibson, eds. *Perspectives on Quine* (Oxford: Blackwell, 1990).

Bilgrami, Akeel. *Belief and Meaning: The Unity and Locality of Mental Content* (Oxford: Blackwell, 1992).

Brandom, Robert. *Making It Explicit: Reasoning, Representing, and Discursive Commitment* (Cambridge, Mass., Harvard University Press, 1994).

Brandom, Robert. "Vocabularies of Pragmatism: Synthesizing Naturalism and Historicism," in *Rorty and His Critics*, ed. Robert Brandom (Oxford: Blackwell, 2000), pp. 156–183.

Brandl, Johannes, and Wolfgang Gombocz, eds. *The Mind of Donald Davidson*. *Grazer Philosophische Studien*, vol. 36 (Amsterdam: Rodopi, 1989).

Cavell, Marcia. *The Psychoanalytic Mind: From Freud to Philosophy* (Cambridge, Mass.: Harvard University Press, 1993).

Dahlbom, Bo, ed. *Dennett and His Critics* (Oxford: Blackwell, 1993).

Davidson, Donald. "Mental Events," in *Experience and Theory*, eds. L. Foster and J. Swanson (Amherst: University of Massachusetts Press, 1970), pp. 79–101. Page references to reprinting in *Essays on Action and Events* (Oxford: Oxford University Press, 1980), pp. ??.

Davidson, Donald. "Paradoxes of Irrationality," in *Freud: A Collection of Critical*

Essays, ed. Richard Wollheim (Garden City, N.Y.: Anchor Books, 1974), pp. 289–305. Reprinted in Davidson 2001.

Davidson, Donald. "Towards a Unified Theory of Meaning and Action," *Grazer Philosophische Studien* 11 (1980): 1–12. Reprinted in Davidson 2004.

Davidson, Donald. *Enquiries into Truth and Interpretation* (Oxford: Oxford University Press, 1984).

Davidson, Donald. "Deception and Division," in *Actions and Events*, eds. Ernest LePore and Brian McLaughlin (Oxford: Blackwell, 1985a), pp. 138–148. Reprinted in Davidson 2001.

Davidson, Donald. "Incoherence and Irrationality," *Dialectica* 39 (1985b): 345–354. Reprinted in Davidson 2001.

Davidson, Donald. "Rational Animals," in *Actions and Events*, eds. Ernest LePore and Brian McLaughlin (Oxford: Blackwell, 1985c), pp. 473–480. Reprinted in Davidson 2001.

Davidson, Donald. "A Coherence Theory of Truth and Knowledge," in *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, ed. Ernest LePore (Oxford: Blackwell, 1986a), pp. 307–319. Reprinted, with "Afterthoughts, 1987," in *Reading Rorty*, ed. Alan Malichowski (Oxford: Blackwell, 1990), pp. 120–138. Reprinted in Davidson 2001.

Davidson, Donald. "A Nice Derangement of Epitaphs," in *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, ed. Ernest LePore (Oxford: Blackwell, 1986b), pp. 433–446.

Davidson, Donald. "Knowing One's Own Mind," *Proceedings of the American Philosophical Association* 60 (January 1987): 441–458. Reprinted in Davidson 2001.

Davidson, Donald. "The Myth of the Subjective," in *Relativism*, ed. Michael Krausz (Bloomington, Ind.,: University of Indiana Press, 1989a), 159–72. Reprinted in Davidson 2001.

Davidson, Donald. "What is Present to the Mind?" in *The Mind of Donald Davidson. Grazer Philosophische Studien*, vol. 36, eds. Johannes Brandl and Wolfgang Gombocz (Amsterdam: Rodopi, 1989b), pp. 3–18. Reprinted in Davidson 2001.

Davidson, Donald. "The Conditions of Thought," in *The Mind of Donald Davidson. Grazer Philosophische Studien*, vol. 36, eds. Johannes Brandl and Wolfgang Gombocz (Amsterdam: Rodopi, 1989c), pp. 193–200.

Davidson, Donald. "The Structure and Content of Truth," *Journal of Philosophy* 87 (1990a): 309–328.

Davidson, Donald. "Meaning, Truth, and Evidence," in *Perspectives on Quine*, eds. Roger Barrett and Roger Gibson (Oxford: Blackwell, 1990b), pp. 68–79.

Davidson, Donald. "Three Varieties of Knowledge," in *A.J. Ayer Memorial Essays*, ed. A. Phillips Griffiths, *Philosophy* suppl. 30 (1991a): 153–66. Reprinted in Davidson 2001.

Davidson, Donald. "Epistemology Externalized," *Dialectica* 45 (1991b): 191–202. Reprinted in Davidson 2001.

Davidson, Donald. "The Second Person," *Midwest Studies in Philosophy* 17 (1992): 255–267. Reprinted in Davidson 2001.

Davidson, Donald. "Thinking Causes," in *Mental Causation*, ed. John Heil (Oxford: Oxford University Press, 1993a), pp. 3–18.

Davidson, Donald. "Replies to Seventeen Essays," in *Reflecting Davidson: Donald Davidson Responding to an International Forum of Philosophers*, ed. Ralf Stoecker (Berlin: De Gruyter, 1993b).

Davidson, Donald. "Could There Be a Science of Rationality?" *International Journal of Philosophical Studies* 3 (1995): 1–16. Reprinted in Davidson 2004.

Davidson, Donald. *Subjective, Intersubjective, Objective* (Oxford: Oxford University Press, 2001).

Davidson, Donald. *Problems of Rationality* (Oxford: Oxford University Press, 2004).

Dennett, Daniel C. *Brainstorms: Philosophical Essays on Mind and Psychology* (Montgomery, Vermont: Bradford Books, 1978).

Dennett, Daniel C. "True Believers: The Intentional Strategy and How it Works," (1981a). Page references to reprinting in Dennett 1987a.

Dennett, Daniel C. "Three Kinds of Intentional Psychology," (1981b). Page references to reprinting in Dennett 1987a.

Dennett, Daniel C. "Making Sense of Ourselves," (1981c). Page references to reprinting in Dennett 1987a.

Dennett, Daniel C. *The Intentional Stance* (Cambridge, Mass.: MIT Press, 1987a).

Dennett, Daniel C. "Instrumentalism Reconsidered," (1987b). Reprinted in Dennett 1987a.

Dennett, Daniel C. "When Frogs (and Others) Make Mistakes," (1987c). Reprinted in Dennett 1987a.

Dennett, Daniel C. "Midterm Examination: Compare and Contrast," (1987d). Reprinted in Dennett 1987a.

Dennett, Daniel C. "Quining Qualia," in *Consciousness in Contemporary Science*, eds. A. Marcel and E. Bisiach (Oxford: Oxford University Press, 1988), pp. 42–77.

Dennett, Daniel C. "Real Patterns," *Journal of Philosophy* 88 (1991a): 27–51.

Dennett, Daniel C. *Consciousness Explained* (New York: Little, Brown, 1991b).

Dennett, Daniel C. "Back From the Drawing Board," in *Dennett and His Critics*, ed. Bo Dahlbom (Oxford: Blackwell, 1993), pp. 203–235.

Dretske, Fred. *Explaining Behaviour: Reasons in a World of Causes* (Cambridge, Mass.: MIT Press, 1988).

Dretske, Fred. *Naturalizing the Mind* (Cambridge, Mass.: MIT Press, 1995).

Evnine, Simon. *Donald Davidson* (Stanford, Cal.: Stanford University Press, 1991).

Farrell, Frank B. *Subjectivity, Realism, and Postmodernism: The Recovery of the World* (Cambridge, UK: Cambridge University Press, 1994).

Fodor, Jerry. *Psychosemantics* (Cambridge, Mass.: MIT Press, 1987).

Fodor, Jerry. *The Elm and the Expert: Mentalese and its Semantics* (Cambridge, Mass.: MIT Press, 1994).

Fodor, Jerry, and Ernest LePore. *Holism: A Shoppers Guide* (Oxford: Blackwell, 1992).

Fodor, Jerry, and Ernest LePore. "Is Radical Interpretation Possible?" in *Reflecting Davidson: Donald Davidson Responding to an International Forum of Philosophers*, ed. Ralf Stoecker (Berlin: De Gruyter, 1993a), pp. 1–23.

Fodor, Jerry, and Ernest LePore. "Is Intentional Ascription Intrinsically Normative?" in *Dennett and His Critics*, ed. Bo Dahlbom (Oxford: Blackwell, 1993b), pp. 70–82.

Føllesdal, Dagfinn. "Meaning and Experience," in *Mind and Language: Wolfson College Lectures 1974*, ed. Samuel Guttenplan (Oxford: Oxford University Press, 1975), pp. 25–44.

Føllesdal, Dagfinn. "Hermeneutics and the Hypothetico-Deductive Method," *Dialectica* 33 (1979): 319–336.

Føllesdal, Dagfinn. "The Status of Rationality Assumptions in Interpretation and in the Explanation of Action," *Dialectica* 36 (1982a): 301–316.

Føllesdal, Dagfinn. "Intentionality and Behaviorism," in *Logic, Methodology, and Philosophy of Science*, vol. 6, ed. L. Jonathan Cohen et al. (Amsterdam: North-Holland, 1982b), pp. 553–569.

Gigerenzer, G. "The Bounded Rationality of Probabilistic Mental Models," in *Rationality: Psychological and Philosophical Perspectives*, eds. K.A. Manktelow and D.E. Over (London: Routledge, 1993), pp. 284–313.

Grandy, Richard. "Reference, Meaning, and Belief," *Journal of Philosophy* 70 (1971): 439–452.

Haack, Susan. *Evidence and Enquiry: Towards Reconstruction in Epistemology* (Oxford: Blackwell, 1993).

Haack, Susan. "Vulgar Pragmatism: An Unedifying Prospect," in *Rorty and Pragmatism*, ed. Herman J. Saatkamp (Nashville, Tenn.: Vanderbilt University Press, 1995), pp. 126–147.

Jackson, Frank. *From Metaphysics to Ethics: A Defence of Conceptual Analysis* (Oxford: Oxford University Press, 2000).

Jennings, Ray. *The Genealogy of Disjunction* (Oxford: Oxford University Press, 1994).

Kim, Jaegwon. *Supervenience and Mind* (Cambridge, UK: Cambridge University Press, 1993).

McDowell, John. *Mind and World* (Cambridge, Mass.: Harvard University Press, 1994)

Malpas, Jeffrey E. *Donald Davidson and the Mirror of Meaning* (Cambridge, UK: Cambridge University Press, 1992).

Nietzsche, Friedrich. *The Will to Power*, trans. Kaufman and Hollingdale (New York: Vintage, 1968a).

Nietzsche, Friedrich. *On the Genealogy of Morals*, trans. Kaufman and Hollingdale (New York: Vintage, 1968b).

Preyer, Gerhard, Frank Siebelt, and Alexander Ulfig, eds. *Language, Mind and Epistemology: On Donald Davidson's Philosophy* (Dordrecht: Kluwer, 1994).

Quine, W.V. *Word and Object* (Cambridge, Mass.: MIT Press, 1960).

Quine, W.V. *Pursuit of Truth*, 2nd edn. (Cambridge, Mass.: Harvard University Press, 1991).

Rorty, Richard. "Introduction," in *The Linguistic Turn*, ed. Richard Rorty (Chicago: University of Chicago Press, 1967), pp. 1–39.

Rorty, Richard. *Philosophy and the Mirror of Nature* (Princeton, N.J.: Princeton University Press, 1979).

Rorty, Richard. *Consequences of Pragmatism* (Minneapolis: University of Minnesota Press, 1982).

Rorty, Richard. "Pragmatism, Davidson, and Truth," in *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, ed. Ernest LePore (Oxford: Blackwell, 1986), pp. 333–368. Reprinted in Rorty 1991a.

Rorty, Richard. "Non-reductive Physicalism," in *Theorie der Subjektivität*, ed. Konrad Cramer et al (Frankfurt: Suhrkamp, 1987), pp. 278–296. Reprinted in Rorty 1991a.

Rorty, Richard. *Contingency, Irony, and Solidarity* (Cambridge, UK: Cambridge University Press, 1989).

Rorty, Richard. *Objectivity, Relativism, and Truth* (Cambridge, UK: Cambridge University Press, 1991a).

Rorty, Richard. "Inquiry as Recontextualization: An Anti-Dualist Account of Interpretation," in *The Interpretive Turn*, eds. James F. Bohman, David R. Hiley, and Richard Shusterman (Ithaca, N.Y.: Cornell University Press, 1991b), pp. 59–80. Reprinted in Rorty 1991a.

Rorty, Richard. "Putnam on Truth," *Philosophy and Phenomenological Research* 52 (1992): 415–418.

Rorty, Richard. "Human Rights, Rationality and Sentimentality," in *On Human Rights: The 1993 Oxford Amnesty Lectures*, eds. Susan Hurley and Stephen Shute (New York: Basic Books, 1993a), pp. 112–134.

Rorty, Richard. "Putnam and the Relativist Menace," *Journal of Philosophy* 90 (1993b): 443–461.

Rorty, Richard. "Holism, Intrinsicality, and the Ambition of Transcendence," in *Dennett and His Critics*, ed. Bo Dahlbom (Oxford: Blackwell, 1993c), pp. 184–202.

Rorty, Richard. "Consciousness, Intentionality, and the Philosophy of Mind," in *The Mind-Body Problem: A Guide to the Current Debate*, eds. Richard Warner and Tadeus Szubka (Oxford: Blackwell, 1994).

Rorty, Richard. "Is Truth a Goal of Inquiry? Davidson versus Wright," *Philosophical Quarterly* 45 (1995a): 281–300. Reprinted in Rorty 1998.

Rorty, Richard. “Philosophy and the Future,” in *Rorty and Pragmatism*, ed. Herman J. Saatkamp (Nashville, Tenn.: Vanderbilt University Press, 1995b), pp. 197–205.

Rorty, Richard. “Reply to Haack,” in *Rorty and Pragmatism*, ed. Herman J. Saatkamp (Nashville, Tenn.: Vanderbilt University Press, 1995c), pp. 148–153.

Rorty, Richard. “McDowell, Davidson, and Spontaneity,” *Philosophy and Phenomenological Research* 58 (1998a): 389–394.

Rorty, Richard. *Truth and Progress* (Cambridge, UK: Cambridge University Press, 1998b).

Rorty, Richard. “Pragmatism as Anti-Authoritarianism,” *Revue Internationale de Philosophie* 53 (1999): 7–20.

Saatkamp, Herman J., ed. *Rorty and Pragmatism* (Nashville, Tenn.: Vanderbilt University Press, 1995).

Stich, Stephen. “Dennett on Intentional Systems,” *Philosophical Topics* 12 (1981): 39–62.

Stich, Stephen. *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation* (Cambridge, Mass.: MIT Press, 1990).

Stich, Stephen. “What is a Theory of Mental Representation?” *Mind* 101 (1991): 243–261

Wilson, N.L. “Substances without Substrata,” *Review of Metaphysics* 12 (1959): 521–539.

Wright, Crispin. *Truth and Objectivity* (Cambridge, Mass.: Harvard University Press, 1992).

Bjørn Ramberg
Professor of Philosophy
Department of Philosophy
University of Oslo
Boks 1072 Blindern
NO-0316 Oslo
Norway