in Stephen Turner and William Outwaite (eds.),
Handbook of Social Science Methodology, Sage Publishing (2007)

16

Rationality and Rationalist Approaches in the Social Sciences

David Henderson

FRAMEWORK

Rationalism may be understood as the philosophical position asserting a certain distinctive epistemic status for certain classes of claims—that asserts or supposes that there are a priori knowable truths. On this understanding, one is a rationalist if one holds that there are certain necessary truths that can be justifiably believed (and that would then count as knowledge) independent of empirical evidence for their truth. This is a somewhat minimalist understanding of rationalism (although these days there are those who would count themselves as rationalist merely by virtue of embracing this much).

Rationalism has traditionally been understood as making a stronger claim. In the modern period, empiricist philosophers such as Locke and Hume sought to debunk what they believed to be the pretensions of rationalist thinkers such as Descartes and Leibniz. The empiricists would have granted there were some claims that satisfied the rather minimal characterization just now associated with rationalism (the claim that some truths were a priori knowable). But, they hastened to add that whatever was so knowable would be something on the order of definitional truths—claims that there were necessarily true as a consequence of the character of, and relations between, the ideas or concepts employed in those claims. All unicorns are mammals—necessarily, since our idea of a unicorn is of a rather particular horsey thing, and our idea/concept of a horse is the idea of a particular sort of mammal (or so the rather plausible story goes). Nothing would count as a horse, and thus as a unicorn, were it not a mammal. But, they insisted, this of itself does not guarantee that there are any unicorns, or horses, in the world. For that, field work, or trips through the country with one's eyes open, would be needed.

While one might know by reflection the relations between our own ideas, said Hume, it is a wholly different question whether there is, in fact, anything in the world satisfying those ideas or concepts. Matters of fact could

only be justifiably believed, could only be known, empirically. In opposition, the rationalists insisted that there were at least some things beyond the creations of our own ideacraft that could be known a priori—they insisted that one could know substantive truths about the world by reflection, without reliance on experiential evidence. Descartes, for example, thought that we could know that we were non-material souls, that God existed, and that material objects were (Euclidean) three-dimensional extended things. (Perhaps there should be a three-strikes rule applied to philosophies.)

In any case, were we to be fully faithful to the terms of this venerable debate, we would need to refine our characterization of rationalism: we would need to understand rationalism as the view that there are certain *substantive* claims that are both necessary and can justifiably be believed (and thus known) independent of empirical evidence for their truth—where a claim is substantive if it is not 'merely definitional', or 'analytic' or guaranteed by the content of its featured concepts.

I mention the traditional and more robust understanding of rationalism only to explicitly lay it to the side. There are multiple reasons for focusing on the minimal understanding in this contribution. Several reasons have to do with the state of play in contemporary philosophy.

First, the idea of a truth that is guaranteed by the semantics of its elements no longer seems to be what it used to be—and the changes have significantly complicated the philosophical landscape. A central development has come with the advent of what is termed 'externalist semantics'. Up until the 1970s, almost all thought about the meaning family (ideas, meanings, intensions, concepts, semantics) supposed that these things were settled by what went on inside a given individual. While such things as meanings or concepts might be abstract entities, whether a given individual entertained or deployed a given meaning or concept in a given stretch of thinking was thought to depend on what occurred within the skin (or perhaps head) of that individual.[1]

This picture was challenged by a line of thought developed by Kripke (1972), Putnam (1975a), and Burge (1979, 1992). They argued that what made for, or constituted, the concepts in play could include elements of the individual's social and physical environment—and was thus not wholly internal to the individual: thus the idea of an externalist semantics). On this view, at least some elements of the semantics of (at least some important) concepts are not (or need not be) accessible to agents employing those concepts. As a result, there could be claims whose truth are guaranteed by the semantics of the concepts featured in them, but which could not be appreciated by those individuals merely by their drawing on whatever makes for an individual's possession of the relevant concepts. Perhaps all a priori knowable truths are conceptual truths, it might then be said, but, if externalism is correct, not all conceptual truths are a priori knowable (even by those who count as conceptual adepts at a given time). Even more significantly, when concepts (semantic entities, meanings, and the like) come to be conceived as rather *more than* ideas in individual heads, the suggestion that a priori truths might turn out to be conceptual truths does not seem as threatening to their significance as was once readily supposed. The empiricist idea that a priori truths might be limited to conceptual truths no longer lends itself to the deflationary rephrasing, '*mere* conceptual truths'.

Second, perhaps influenced by such considerations, those with avowedly rationalist inclinations have comes to think largely in terms of conceptual truths without feeling insignificant (for example, Bealer, 1987; Chalmers, 1998, 2002a, 2002b; Jackson, 1998; Peacocke, 1992). Conceptual truths pack some punch, at least if contemporary understandings are roughly correct.

Now that conceptual truths have come to be thought of as 'more muscular' or substantive as a class, the fan of a priori knowledge has come to face a new challenge: to explain how it can be that those who are relatively proficient with the concept have, by virtue of that conceptual competence, access to powerful

elements of that semantics. This has been a matter of identifying a component of the semantics of the concept that are accessible at least to those who count as 'possessors' of the concepts involved.[2] But, we will not now detail the lines of the contemporary debate over the epistemology of the a priori.

These philosophical preliminaries do serve to indicate why it is that contemporary rationalists do not seem much concerned with what their modern ancestors would have thought crucial—why they commonly are not much concerned to show that there are a priori knowable truths that are not 'merely conceptual truths'. They also serve to explain to my readers why my discussion of 'rationalist approaches to the social sciences' will focus on positions regarding the subjects of social-scientific thought (on positions regarding beliefs, thought, actions, and the like) that might be thought to be conceptual in their foundation.

In connection with the social sciences, the central matter on which a priori truths have been sought or sensed by those with rationalist inclinations has been the role of rationality in the explanation of action. Put starkly and overly simply, it is said to be a priori that finding one's subjects to have beliefs and desires that makes their actions rational explains their actions, while failing to do so leaves their action unintelligible and unexplained. This is said to follow from the concept of an action—which is said to involve the idea of a behavior engendered by reasons. The concept of a reason is said to involve the idea of a contentful state that bears a (normatively approvable, i.e., rational) support relation to some other contentful state (that for which it is a reason). Putting these thoughts together, it is said to be a conceptual truth that actions are the sort of thing engendered by contentful states (prominently, beliefs and desires) that rationally support the decision to undertake that action.

Doubtless, such points would need to be sharpened and qualified in certain ways (I will try for more nuance below). At this juncture, it is useful to state the general thrust of the rationalist position. One then sees that the putative a priori truth—that actions have rational antecedents and explanations, or something along this line—would both inform and constrain work in the social sciences. It would inform a kind of explanatory practice—and one apparently 'on the cheap'. After all, one would not have to develop well-evidenced generalizations or descriptive theories of cognition; one would not need empirical theories of human cognition that underwrite the explanatory practice in question. Instead, one's own normative principles of reasoning would supposedly turn the trick of informing and supporting explanations. One's own normative principles or reasoning competence, representing or tracking support relations between contents, would structure explanations underwritten by the very concepts of action and reasons. Such normative principles would also be constraining (again a priori constraining)—normative principles would need to figure in explanation, thereby limiting the kind and character of explanation one could employ in the human sciences. Such, in outline, is the central rationalist approach to the social sciences. In all variations of this generic approach, some significant degree of rationality in belief, desire and action is thought to be a priori conceptually guaranteed.

THE RATIONALIST BRIEF

Why embrace some variant on the basic line of thought sketched above? In surveying the rationalist case, one can begin by reflecting on the everyday practice of explaining an action by attributing reasons to the agent or agents. This is pretty pervasive. One encounters it in discussions of friends, enemies and acquaintances. One encounters it in histories, in newspapers, in meetings and just about anywhere that actions are up for explanation. It is reasonably taken to reflect some of our must fundamental understandings of what it is to act—and very central elements of this understanding might reasonably be taken to reflect our concept of an action,[3] and so to be grist for the rationalist

mill. What then does the rationalist find when reflecting on the practice of explaining actions by citing reasons in the form of beliefs, desires, and the like?

When explaining an action by citing some of an agent's beliefs and desires, it seems important that the (typically small) constellation of the agent's contentful states mentioned be such as would count as a reason for so acting. If the cited beliefs and desires seem unrelated to the apparent action, then no explanation has yet been provided. To say that an agent had such and such a reason for some action seems to require that the beliefs and/or desires here mentioned *rationally favor* that action. This is *not* to say that they need make the action rational for the agent *all things considered* (that is, given the full range of the agent's beliefs and desires). It is to say that, considered just of themselves, they must rationally count in favor of doing such an action. This observation is crucial for rationalists, and is taken to reveal the most fundamental outlines of the coordinate concepts of an *action* and a *reason*. An action is the sort of thing undertaken for a reason. A reason for an action is a contentful state, or perhaps small constellation of such states, that presents that action, that undertaking, as rationally a good thing. Because this much only supposes that the reasons cited make the action narrowly rational, rather than all things considered rational, the commonsensical practice of explaining by citing reasons may be termed narrow rationalizing explanation.

Donald Davidson is perhaps the most influential of contemporary commentators on rationalizing explanation, and the rationalist brief here recounted is indebted to his writings. He argues that rationalizing explanation in terms of reasons turns on the idea of mental states as *rational causes*. Thus, it is taken to be central to our concept of an action, and to our most rudimentary explanatory practice with respect to actions, that actions are events with (in at least some minimal sense) a rational cause. To explain an action of an agent, the beliefs and desires attributed to the agent must 'make the act intelligible', must amount to rational reasons for so acting. Indeed, Davidson (1982: 299) insists that such rationalizing explanation provides 'the only clear pattern of explanation that applies to action'. It is a pattern that has been elaborated in significant ways within economics, for example. But, whatever we ultimately want to make of these elaborations (and that is something to be considered below), the core rationalist idea seems to be that this 'only clear pattern'—this practice of narrow rationalizing explanation—represents a deep element of our concepts of *action* and *reason*. This is to say that *action* and *reason* are coordinate concepts—concepts made for each other—and that it is then a priori necessary that actions be rational at some significant level.

Suppose that an individual agent undertakes some action that can be given a relatively uncontroversial characterization. Commonly, there will be several constellations of the agent's beliefs and desires that could have prompted such an action. That is, within the agent's full array of beliefs and desires, there may be several small constellations that are each made up of beliefs and desires so related as to present that action (under the uncontroversial understanding) as a good thing. Understanding and explaining the action will then depend on identifying what—from among the various reasons that the agent might have had (even did have) for so acting—was *the reason* that the agent did the action. For example, the agent may have given a gift. Is that to be explained by a strong altruistic desire to meet a need of the receiver, or by a desire to be remembered in an upcoming decision process (or in a will), or by a nagging guilt over the giver's own role in a recent decision, or by some combination of these? Perhaps the agent had all these reasons for giving a gift.

To explain the action—the giving of a gift—one must appreciate which reasons were operative or dominant. To correctly explain the action, one must be able to say that the agent did the action *because* the agent desired such and such, and believed this and that, and sometimes this requires saying that the agent did the action for these

reasons rather than because of certain other constellations of beliefs and desires that the agent might (even did) have. As Donald Davidson (1980a) argued, there is really no account to be given of the 'because' here, without invoking a causal relation. It will not do to say that the beliefs and desires *simply* 'make intelligible' the action—for any of various constellations would do that—and of only one (more or less sprawling) constellation of the agent's actual reasons will it be correct to say that the agent did the action because of those reasons. Only prominent members of this constellation will do as explanations (or will serve as the core of an explanation) of the action in question. So, the rationalist insists, it is a priori that actions are caused by constellations of beliefs and desires that together rationally indicate the desirability of the action. Not only must the agent have reasons for his or her action, but some constellation of such reasons must have been central to the processes that causally produced that undertaking.

Many a piece of behavior—say, the giving of a gift—might be motivated by various beliefs and desires. Indeed, the piece of behavior might then be understood as one of several actions. A given stretch of behavior might be understood as a selfless gift, an obsequious bit of pandering, a response to felt guilt, and so on. In important respects, the character of the action, as action, itself depends on the reasons that caused it. (This is so, even though we clearly allow some latitude for various descriptions of a given action, and some of these do not suppose that the agent was motivated to undertake an action so described. For example, we allow actions to be described in terms of their unintended consequences. So the present point would need to be understood in a fashion consistent with the possibility of such re-descriptions.) Insofar as the character of an action depends on the engendering reasons, insofar as this much is a deep element of the concept of an action, it is conceptually mandated that actions are caused by reasons—by contentful states that rationally (perhaps narrowly rationally) present them in a favorable light.

The upshot of these few quick paragraphs, focusing on the coordinate concepts of *beliefs* and *desires* as *reasons*, and of *actions* as caused by reasons, has been to present the most rudimentary prima facie case for the rationalist idea that it is a priori that actions and reasons must be understood in terms of processes that are preponderantly rational, at least narrowly so. Thus, subject to qualifications, we can announce a general, if crude, result:

Restricted Rationality Principle Concerning Actions (RRP–a)

Actions are behaviors with rational causes. To be an action is to be a bit of behavior which is (at least narrowly) rationalized by a constellation of beliefs and desires that cause it.

The parenthetical qualification is important. For the above paragraphs have not established or defended any sweeping rationality principle—no principle to the effect that actions are rational in virtue of the full set of the agent's beliefs and desires, or that the agents' beliefs and desires are themselves preponderantly rational. (The extent to which one might reasonably seek to extend these general lines of thought to provide for such a stronger rationalist principle will be the concern in much of the discussion below.) In its limited form, it is plausible that RRP represents something central to, something guaranteed by the concepts of *action, belief,* and *desire.*

RRP as depicted above focuses on *actions*—saying that they are behaviors with rational causes. (Thus, I designated it the Restricted Rationality Principle Concerning Actions.) But parallel lines of thought suggest a parallel limited rationality principle with application to beliefs and desires. The result would be conceptually grounded rationality principles that apply to any given belief (RRP–b) and to any given desire (RRP–d)—indicating that these must also be at least narrowly rational in their interrelations with generating and sustaining beliefs and desires. Of course, beliefs and desires are not undertakings, while actions are. So the

arguments supporting these additional rationality principles will need to be somewhat different. Still, it seems central to beliefs and desires that they must be such as could be rendered intelligible in terms of certain of the agent's (narrowly) associated beliefs and desires. The cumulative effect is an understanding of the various items—beliefs and desires, as well as actions—according which these are generated and sustained in a web of contentful supports that are (at least narrowly) rational.

The most parallel line of thought here would take the form of reflection on how one explains a person's having a belief that that person apparently holds (or a desire that an agent apparently possesses). A common explanatory move here is to show that the belief was (fairly obviously) rationally indicated by at least some set of the agent's (salient) other beliefs. Similarly, one commonly explains an agent's desire for some state or object by appeal to apparent antecedent desires salient to the agent, along with beliefs that seem rationally to jointly indicate that the state or object would facilitate the satisfaction of these antecedent desires. Here again, one explains by noticing the agent's reasons for so believing (or desiring). But, here, one need not suppose that the belief (or desire) so explained is rationally *chosen*, as this would involve an implausible voluntarism with respect to belief and desire. Nevertheless, reasons are readily in question when explaining beliefs and desires—and showing the belief (or desire) in question to be reasonable makes the agent and the belief (or desire) 'intelligible'. Here also, one only succeeds in explaining the agent's holding the belief (or desire) in question when the antecedent beliefs (and desires) alluded to are (or feature in) 'the reasons' that the agent believes (or desires) as he or she does. Again, causal dependencies are at issue.

Generalizing we arrive at a restricted rationality principle with respect to belief and desire:

RRP–b/d: Beliefs and de\sires are states that interact in significantly contentfully appropriate ways.

But, again, as with RRP–a, it is important to keep in mind that the rationality indicated in these lines of thought should be understood as narrow rationality: a matter of not particularly subtle rationality involving some limited range of beliefs and desires that are salient to the agent in those processes that issue in the beliefs, desires and actions being explained. There does not yet seem to be a *conceptual* demand for a subtle and far-reaching holistic rationality involving vast ranges of a given individual's beliefs and desires and turning on sophisticated support relations.

A related argument will begin to push one towards the stronger principles that most rationalists regarding rationality in the human and social sciences would also embrace. It comes into view when pursuing a question only touched upon above: how can one determine which of the possible beliefs and desires that might rationalize an action are 'the agent's reasons for' so acting? This is best thought of as a complex question, one that can be decomposed into at least two questions, each of which might motivate rationalist argument.

One question is: which of the beliefs and desires that might help to rationalize the action are possessed by the agent? Action is the joint product of interacting beliefs and desires, as already noted. In paradigmatic cases, one can take an action as reflecting a choice, and look for sets of beliefs and desires that might have (rationally) produced it. But which of these are possessed by the agent? Further, this way of putting the question obscures the magnitude of the interpretive task, for the action undertaken itself can only be appreciated or understood against an understanding of the beliefs and desires engendering it. Thus, as Davidson (1984c, 1980c) argued, there seem to be *three unknowns* that must be sorted out on the basis of the behaviors presented to an observer/interpreter: beliefs, desires and actions/intentions. Here we confront directly what has been termed *charity in interpretation*. *If we presuppose rationality* on the part of the agent, information about any two of

the three (beliefs, desires, actions undertaken) will allow one to determine the third. Given any two of the three, and the presupposition of rationality, one can 'solve for the other'. For example, desires can be determined from actions undertaken and from beliefs. There seems to be no alternative but to suppose some measure of rationality, or so it is argued. But, says Davidson, one must presume more.

The rationality envisioned here as presupposed or imposed as a part of the interpretive endeavor seems to take one beyond the narrow rationality at issue in the above discussion. Significant light is shed on an agent's 'standing' beliefs and desires by looking at a wider set of actions undertaken by the agent in question. Of course, there might be various interpretations that one might plausibly put on these—but the idea is that one might constrain workable alternative interpretations by looking for plausible *coherence over the standing beliefs and desires* attributed against *the 'data points' provided by large sets of behaviors*. But, as one imposes a rationality assumption over wider sets of actions—and their precipitating reasons—one imposes a *more holistic* and less narrow sort of rationality in interpretation.

Focus on this issue: how can one get some principled grasp on any one pair—for example, beliefs and actions—within the triad of unknowns? Again, as Davidson would insist, there is in play a charitable presumption having to do specifically with beliefs—and beliefs provide a common entering wedge for interpretation. Davidson (1980c, 1984b) holds that we can and must charitably provide some determinacy to the belief element in the triad by assuming that most beliefs about any given subject are true. (For now, I am looking to Davidson's understanding of charity in interpretation for inspiration in developing a rationalist brief. Later, I will compare this understanding with a more cautious or limited charitable approach is suggested in Quine's writings.) Davidson's charitable policy with respect to the attribution of beliefs serves as a way of 'holding belief constant enough' to get on with the business of interpretation—in order to solve for other factors in the interpretive trinity, where solving for the 'other factors' involves charitable presumptions regarding rationality**.** Rationalists, of course, insist that that charitable presumption of rationality is conceptually mandated—a conceptually mandated principle significantly stronger than RRP.

Thus, to tentatively summarize the rationalist response to the first question: to figure out what beliefs and desires a subject holds, we are instructed to make two charitable presumptions in interpretation. We are to presume some significant degree of correctness of belief on the part of the subject, at least with respect to some significant range of matters. Further, we are to presume some significant measure of rationality in the structure and interaction of beliefs and desires—so that the content of a belief or desire will be understood in terms of its patterned relation to other beliefs and desires, given their contents. To cut down on alternative sets of attributable beliefs and desires, we are to extend the range of observed choices/undertakings and speech to be accounted for.[4] This will allow one to rule out some alternatives explanations for a given action, as the eliminated explanation would be inconsistent with the agent's pattern of actions and motivations. The crucial point is that in casting our interpretive and explanatory net more broadly, one must apparently presume *stronger and more holistic forms of rationality* in beliefs, desires and decisions.

Then there is the second question: which of the constellations of rationalizing reasons to be found in the agent represents 'the reason' that the agent so acted? Here, the rationalist idea is that, in keeping with what has been said to this point, the determination of an agent's reasons for acting in a certain fashion (that is, for determining what reasons were *the* reasons, for determining which were causally salient) is a matter to be settled by a kind of inference to the best explanation, constrained by the range of that agent's beliefs and desires (charitably determined) and by

the kind of holistic rationality necessarily supposed in their determination.

The upshot of all this might now be formulated as an unrestricted (or much less restricted) Rationality Principle in two clauses.

Rationality Principle (RP)

RP–a: Actions or undertakings are behaviors with rational causes—beliefs and desires cause them, and do so by virtue of certain contents that make them reasonable, significantly holistically rational, and thus intelligible.

RP–b/d: Beliefs and desires (while not choices or voluntary in the same limited sense in which actions/undertakings might be) are caused products of each other, in a fashion that reflects contents and makes for their being reasonable, significantly holistically rational, and thus intelligible.

Much thought in economics and related disciplines is aptly understood as continuous with the common thought about beliefs, desires and actions that we have been considering. In various ways, economists have sought to develop precise mathematical ways of thinking about choice, ways that attempt to measure strength of desire and degree of belief in scales with understood properties. While such thought may be more articulate and more careful than everyday talk of beliefs, desires and actions, it must be recognized that something like RRP and RP plays a parallel role there. The central issue for our purposes is whether, or to what extent, something like RP has an a priori status there.

This will need to serve as the main brief for rationalist approaches to the social sciences—particularly with regard to the putative a priori role of rationality in understanding action. It reflects the considerable reasons that one might have for advancing several claims as a priori in character:

1. That actions are behaviors with rational causes—at least the agent must have certain belief states and desire states that make it narrowly rational for the agent to undertake that action (RRP–a).
2. That beliefs and desires are states which interact in the ways that are significantly rational—and perhaps that they are states with the relevant content by virtue of this pattern of dependencies. Here, it seems, rationality is partially constitutive of beliefs and desires. Again, we have the idea of a cause that is also a reason (RRP–b/d).
3. To exhibit or reveal the narrow rationality of a choice or action is to explain it. To exhibit the significant rationality of an agent's beliefs or desires is to explain them. In both cases, one is exhibiting the choice or mental state to be caused by its rational antecedents in 'intelligible' or 'reasonable' ways.
4. Overall, to qualify as a belief, desire or action, a result requires more than the narrow rationality of choices or actions undertaken, or the narrow rationality of belief and action—as it requires that the choice and its belief and desire parameters themselves exhibit rationality of some significant holistic or extended sort (although certainly not perfect holistic rationality) (RP).

Alexander Rosenberg (1985, 1988) provides a particularly clear and striking way of advancing the rationalist points made here.[5] He takes note of the ways in which finding rationality is taken to be explanatory—by making intelligible a choice or undertaking by revealing the reasons that motivate it. He is led to give expression to RRP when he formulates an 'oversimplified general statement [that] seems to lie behind ordinary explanations of human action' (1988: 25):

> [L] Given any person *x*, if *x* wants *d* and *x* believes that *a* is a means to attain *d*, under the circumstances, then *x* does *a*.

Taken as expressing a narrow form of rationality, the exhibition of which is central to intentional explanation, [L] serves to express the sort of thin putatively a priori claim—the RRP—envisioned by rationalists.

Then, reflecting on how beliefs and desires are thought to conspire holistically to produce a choice, Rosenberg is led to suggest a more full-bodied principle representing how rationality is putatively involved in the explanation of action and in the interpretation of agents' beliefs and desires. The suggestion is that something very like normative decision theory represents a (again fairly commonsensical) refinement on [L]—call it

[L^r]. It is worth noting that these refinements are developed and advanced largely 'from the armchair'. That is, the wrinkles associated with decision theory can certainly seem 'natural' when systematically reflecting on sorts of cases in which the antecedent of [L] would be satisfied and yet the agent not undertake the indicated action, one is likely to think: 'Now, *x* may want *d* and *x* believe that *a* is a means to attain *d*, under the circumstances. But what if *x* wants *g* more than *d* and believes that getting *g* is incompatible with doing *a*.' Just as [L] expresses a weak and thin rationality principle thought to a priori hold action, so decision theory would constitute a highly substantive refinement—[L^r]—and is taken by Rosenberg to express a much more constraining and substantive, putatively a priori claim: in effect, RP.

According to Rosenberg (1988: 30–6), you should come to a striking realization: little testing and refinement of [L^r] is possible. The reason is again rooted in the principle of charity in interpretation: [L^r] is supposed in arriving at the interpretations that would be the necessary preliminaries to determining whether agents conform to [L^r]. Rosenberg concludes that a rationality principle along the lines of [L^r] functions as something like a 'definition', rather than providing an empirically testable or refineable description of cognitive tendencies or as a nomic generalization (1988: 33).

CRITICAL EVALUATION, STAGE ONE: REGARDING RRP

Let us begin by looking at the least demanding element of the position: the Restricted Rationality Principles. We can focus on RRP–a. We should keep in mind two respects in which RRP–a advances a very limited claim. The first has to do with the etiology of undertakings or actions; it holds only that, among the agent's vast set of beliefs and desires, there are or were some contextually salient beliefs and desires that both (a) featured in the near causal antecedents of the choice or undertaking and (b) added up to a reason for so acting (jointly portraying the action as good to do). It does not claim that these causal antecedents, these contextually salient occurrent beliefs and desires, are themselves ultimately rational for the agent to hold—ones that makes rational sense in light of the agent's wider set of beliefs and desires, or ones that result from rational inquiry or deliberation. Second, it does not hold that the belief is presented in an overall favorable light by the *total set* of the agent's beliefs and desires—or even by a very extensive set. It does not suppose that the action is 'all things considered' rational, given the agent's full range of beliefs and desires.

This said, it becomes plausible that RRP–a (and Rosenberg's [L]) might indeed be a central element of the concept of an action. Paradigm cases of actions are intentional behaviors undertaken for certain reasons—such reasons jointly amount to a representation of that undertaking as a way of attaining certain of the agent's ends. This does not mean that every action is intentional—that (so described) it was intended by the agent. There are familiar ways of describing an action that do not turn on the agent's understandings or representation of that undertaking. For example, we sometimes describe an action in terms of consequences that were not intended. Descriptions in terms of institutional consequences can be a case in point. It is said that in 2000 a significant number of voters in Florida voted for the Republican candidate for president unintentionally. Yet such agents possessed reasons for their undertaking *as they understood it.* As they understood their action or undertaking, they had their reasons, and these made it out to be desirable in the sense envisioned in RRP–a.

Further, RRP–a should be understood to apply to less paradigmatic cases of actions or undertakings. In impulsive acts which were not conditioned by significant deliberation, one might 'just have felt that it would be nice to' do such-and-such (pinch the child, crack a joke, run to the top of the hill …). Yet, there is a sense in which the agent inarticulately

'chose' to engage in an action which 'seemed good at the time'. Here, 'the constellation of beliefs and desires that cause' the action may be rather thin, but they present it in a good (if feeble) light, and this conforms to RRP–a.

I do find it plausible that this much is conceptually guaranteed by the very concept of an action. Were we to give up on the notion that this holds for a wide range of those events that we have been thinking about as undertakings, we would thereby have compelling reason to give up on the idea that there are any undertakings and any actions at all.

RRP does not provide much guidance in settling the real substantive questions that concern social scientists, or even those that concern folk in everyday contexts. It does not, for example, give much direction for determining what are 'the agent's reasons' for a given stretch of behavior that might well be an action. For many episodes that one plausibly treats as some undertaking or action, there may be various constellations of beliefs and desires that the agent might hold, and that would put the undertaking (under some interpretations) in a favorable light, and it is a significant question which of these sets the agent had, which were the agent's reason for so acting, and what then is the intentional character of this undertaking. *If there is to be strong a priori guidance or constraint on the explanation of actions, there would need to be markedly stronger a priori principles.*

EVALUATION, STAGE TWO: REGARDING RP

Empirical Resources and the Revisability of Rationality Expectations

We need to understand how the range of an individual's beliefs and desires interact as considerations and compose themselves so as to yield a choice. We need to understand how various reasons or considerations resolve themselves into something that might be termed *the* reason—a causal vector of meaningful states in which some stand out as dominant or controlling of the subsequent course of action.

The central idea in the more robustly rationalist approach to the human sciences, and to the place of rationality in those sciences, turns on one simple idea that is said to be guaranteed by the coordinate concepts of *belief, desire* and *action/undertaking*: *that beliefs and desires as considerations interact according to holistic rational principles, and thereby compose themselves, all things considered, into rational choices—choices that are, from the point of view of such considerations, a rational resolution*. As reflected in Davidson's and Rosenberg's writings, the projected a priori rationality in action is of the sort represented by normative decision-theory (and logic, and epistemology). It is thought to be conceptually necessary that, in the preponderance of cases, actions undertaken are the rational product of the strengths of the agent's various desires and of the agent's beliefs concerning the propensity of various course of action to produce or frustrate those desires. The rational course of action for an agent is that with the highest expected value among those courses of action open to the agent—where the expected value of an action is understood as the sum of the possible (positive and negative) outcomes of that course of action (as conceived against the agent's background beliefs), each weighted by the agent's understanding of the probability of that outcome given that course of action.

Such is the full-blooded conception of rationality *in action* that the proponent of RP commonly envisions. This understanding itself supposes an understanding of a corresponding holistic rationality in belief and desire. In keeping with the principle of charity, agents are understood as possessing a rich set of standing beliefs and desires. These may evolve over time, under prompting by experience and reflection. But, these standing states are understood to be 'reasonably' constant, and changes in them are thought to

be of a largely rational character. Their interactions or interrelations are said to be such as to evince a preponderance of rationality**.** Rationalists seem less articulate on the precise character of the rationality that is supposed to be guaranteed here—and they resort to general and hedged formulations. Davidson writes of a 'large degree of consistency' (1980b: 221), and of significant conformity with 'stipulated structures' of a normative character (1980c: 6–7), of 'imposing our logic' in interpretation. 'It is uncertain to what extent these principles can be made definite—it is a problem of rationalizing and codifying our epistemology,' says Davidson (1980c: 7). It should now be clear that RP amounts to a rather significant rationality claim regarding both cognition and action—a claim taken to hold a priori of all creatures with beliefs and desires, creatures who undertake actions. Compared to the first small rationalist step (RRP), this second step (RP) seems quite a stretch! In keeping with the conceptual status claimed for it, it is said to so constrain both the attribution and explanation of actions and cognitive states that it is neither at risk of significant empirical challenge nor susceptible to significant empirical refinement. Call this the *strong rationalist position* regarding rationality in the human sciences.

Confronted with such sweeping claims derived from abstract philosophical reflection, one does well to approach them with caution. If strong rationalism is correct here, then something along the lines of full normative decision theory descriptively applied—[L^r]—must be correct. If strong rationalism is correct, then [L^r] should not be subject to empirical test or refinement. [L^r] could not be subjected to empirical refinement or test because [L^r] would play a conceptually grounded constraining role in the attribution of beliefs and desires; attributions involving significant violations of [L^r] would count as problematic (indeed as conceptually incoherent) interpretations.

However, there is reason to believe that [L^r]—in effect, decision theory deployed as a descriptive account of human cognition[6]—*is* subject to empirical test, i.e., [L^r] can be empirically shown inadequate and refined. The best reason for thinking so is that there is reason to think that [L^r] *has been* tested and found inadequate—prompting empirical refinements. An apparently instructive example can be found in the well-known work of Tversky and Kahneman (1974). Tversky (1975) contrives experimental situations in which people's responses give us empirical reasons for revising our understanding of human cognitive tendencies—evidence indicating that [L^r] must be abandoned or, what amounts to the same thing, significantly revised. Consider a set of situations and results that Tversky discusses. The situations are of a common sort found in studies of decision-making under uncertainty: choices between gambles. (Using the standard notion, (X,P,Y) will represent a gamble where one will receive X with a probability of P, or Y with a probability of $1 - P$.) Tversky presented subjects with a choice between gambles A and B:

A= ($1000, 1/2, 0), B= ($400)

Presented with this choice, almost all subjects prefer the 'sure thing', B. They do this despite the fact that A has a greater actuarial value: $500.

Such results are not themselves news within standard decision theory, and present no immediate threat to [L^r]. After all, it is common to distinguish between the amount of goods or money to be had and its 'utility'. The latter is conceived as a subjective, nonlinear, function of the former. The common postulation of decreasing marginal utility—a concave positive utility curve—is clearly enough to accommodate the results obtained in connection with choice situations of just this first sort. One need only claim that, commonly, $u(\$400) > 1/2u(\$1000)$. This response is just what the strong rationalist would anticipate: [L^r] is not impugned by the above results because we interpret our subjects on the basis of its charitable insistence on standard normative decision theory.

However, this is an overly simple description of our interpretive practice. At some point, and Tversky's work takes us to such a point, the [L^r]-informed identification of values held by subjects comes to clash with other constraints—and [L^r] can give way. This begins to be in evidence in connection with a second choice situation, one produced by multiplying the probabilities of gains by 1/5. That is, subjects are presented the choice between C and D:

C= ($1000, 1/10, 0), D= ($400, 1/5, 0)

If the explanation of the choices found in the first situation were really the concave shape of the subjects' preference curves, then we could expect a preference of D over C. However, that is not what is observed. Within the confines of standard decision theory, the overall pattern of choices is 'incompatible with any utility function' (Tversky, 1975: 166). Tversky's results suggest that there is a '*positive certainty effect* ... [in which] the utility of a positive outcome appears greater when it is certain than when it is embedded in a gamble' (1975: 166). He also provides evidence for a negative certainty effect. Such interactions of utility and probability violates aspects of standard normative decision theory, where it is supposed that there are utility functions (unique up to a certain transformation) associated with particular goods and that such utility functions interact simply with subjective probabilities according to the rule: $u(x)p(x)$.

One tempting response would be to insist that Tversky's subjects just did not understand the situations in the way he supposes. This would be to invoke the strong rationalist position regarding charity and [L^r]. However, and this is crucial, one can raise and address this issue in a principled fashion—one that itself seems empirically informed. Consider, just what was it about the situations that Tversky's subjects plausibly understood differently? They were American college students. Is it plausible that they did not understand talk of 'dollars'? Of that they did not understand the rudimentary mathematical relations between 1000 and 400—or that whatever could be purchased with $400 can typically be purchased in a matched pair with $200 remaining from $1000? The reason that such differences in understanding are not plausible is that in addition to some expectations for certain forms of rationality, we also have expectations regarding roughly when people learn rudimentary matters of importance within their society. We expect such elementary math and monetary units to be learned much earlier than college. Such relatively mundane, but nevertheless empirical, expectations effectively block positing significantly different understandings of the relevant aspects of the situations Tversky presents to his subjects.

It is more plausible that Tversky's subjects understood the probabilities stipulated in ways differing from Tversky's (and ours). And it certainly is true that they may not have developed any sophisticated understanding of probability. But, Tversky's results do not require sophisticated understandings. It seems quite likely that his subjects could have applied talk of probabilities to matters such as coin tosses, urns with colored balls, and whether their car would start next time tried. That would be enough to make Tversky's results telling. (If it was lacking in most people's thoughts, normative decision theory is likely in trouble anyway.) Again, we find some relatively mundane and empirical expectations constraining interpretation—and these could be given further empirical development.

Thus, in addition to some [L]-like expectations serving as constraints on interpretation, we find various more or less mundane, more or less empirically developed, expectations also constraining interpretation—with the result that suggested refinements of [L] can be put under significant empirical test in ways reflected in Tversky and Kahneman's work (to name just one prominent example). Call these empirical constraints—*empirical expectations* (or EE). These EE are a diverse lot. Some are fairly general in character—for example, they may have to do with the power

of human cognitive abilities, whether it is reasonable to expect that someone would 'put certain things together' and appreciate certain implications, with whether it is likely that certain learning or experiences would be recalled from memory, with certain commonality in human motivation, 'needs' and the like. They may have to do with various domains of human cultural phenomena: religion, group identity, political phenomena, the flow of information within various groups, economic phenomena and the like. They may have to do with particular cultures or groups, as in what things are learned when within a certain culture. The point is that such diverse EE provide a significant constraint on our understanding of people, and can make it empirically plausible that one has encountered a case where some proposed development on [L], such as [L^r], is violated. With systematic enough violation, one can have empirical basis for abandoning some proposed development on [L] in favor of others.[7]

The essential issue in evaluating strong rationalism is whether [L^r] plays such a decisive and dominant role in informing what beliefs and desires are attributable to agents that [L^r] is itself rendered immune to empirical pressure and revision. We have just considered a kind of empirical inquiry in which it seems that significant basis is provided for revising [L] in ways that amount to abandoning [L^r]. The suggestion has been that there are multiple empirical constraints on interpretation—EE—that can provide leverage for abandoning or deeply revising [L^r].

Tversky and Kahneman (1974) then advance an alternative to [L^r]—an alternative descriptive account of human choice behavior—which they term prospect theory. It counts as an empirical refinement for several reasons. First, the motivation for abandoning [L^r] in favor of some alternative is empirical. Second, the particular alternative is judged promising and worthy of further empirical investigation because it accommodates the observations obtained in Tversky and Kahneman's work. Third, PT not only accommodates Tversky and Kahneman's observations, but those observations are highly plausible; it seems, in fact, that they are most plausible, given the sorts of empirical constraints in question.

Those of a strong rationalist bent may have conceived of a rejoinder. They will note that the various above-mentioned concrete empirical constraints themselves turn on antecedent interpretations of human beings generally, and of those in more narrow populations such as those from which Tversky and Kahneman's subjects are drawn. After all, how do we know that most folk of college age were long ago exposed to information regarding certain topics? How do we know that humans are capable of learning what little math is needed to recognize the points of significance? These empirical constraints seem to be ploddingly obvious generalizations arrived at on the basis with everyday experience with those very populations, or presumably similar populations. As such, they depend on antecedent interpretation. The rationalist would insist that such interpretation must have been constrained and informed by something like [L^r] all along. It then seems that, in retaining these interpretations—and in making the revisions that constitute PT, or something on this order—one must be, or should be, seeking to diverge from [L^r] in the most minimal fashion. This is to say that [L^r] cannot be empirically revised *much*, and that [L^r] itself serves as a irrevisable constraint from which deviations under interpretation must be minimized. The strong rationalist point might be put in terms of the performance/competence distinction: one may need to attribute moments of irrationality, it is conceded, but these must always be isolated enough to count as mere performance errors against a background of rational competence.

There are reasons for doubting that the rationalist has things quite right here. The essential issue has now to do with the character of the ultimate constraints on interpretation. Is it really the case, a priori, that any empirical refinement of [L^r] would need to

rely on interpretations that largely confirm [L^t]? The rationalist rejoinder requires that *some rather powerful normative model of rationality* (something like standard decision theory together with some parallel account of epistemic rationality) *serves as an invariant constraint on interpretation—that there is some such constraint on interpretation, which is invariant, does not evolve, being set a priori*. Again, in philosophy, this view is prominently associated with Donald Davidson's writings.

The Principle of Charity vs. The Principle of Explicability (Sub-heading level 2)

The principle of charity in interpretation is roughly that one must so interpret as to find those interpreted to be preponderantly rational and believers of mostly truths. (This formulation reflects Davidson's influential development of the principle.) Let us focus on the idea that we must find rationality under interpretation. Here it is crucial to distinguish between two understandings of this supposed constraint on adequate or acceptable interpretation. One sees the first as absolutely fundamental, and the other as derivative and plastic in certain respects. The strong rationalist idea is not merely that we must find our subjects to be reasoning in certain ways, and that many of those ways happen to be rational ways to think, so that we need to find such rationality under interpretation. (That much is congenial to one who sees the constraints on interpretation as a matter of deploying expectations for human reasoning that are commonsensical but ultimately empirical in character.) Rather, the strong rationalist idea is that that the need to attribute rationality is fundamental, that rational ways of thinking (and acting) must—*because they are rational* ways of thinking and acting—be supposed in interpretation. It is the idea that it is by virtue of being rational that certain ways of thinking must be found under acceptable or adequate interpretation. Not all understandings of the principle of charity turn out to suppose this, but strong rationalist understandings (such as Davidson's) do.

The rationalist understanding of the principle of charity contrasts with Quine's understanding of that principle, according to which the charitable constraint is derived and plastic. Quine (who is writing somewhat narrowly of translation) says that we must translate others so as to preserve 'the obvious'—where what is obvious is a matter of empirical psychology. Before discussing attributions of rationality and irrationality, consider the implications of Quine's admonition for attributions of true and false beliefs. Some truths are (in context) relatively obvious—for example, given a context of good illumination where one's subject is at arms length, it should be obvious that one is faced with a rabbit (and not a grizzly bear). (Of course, there might be less frequent contexts involving good light and proximity where it would be at least equally obvious that one is faced with a grizzly bear, not a rabbit.) So, if one's scheme for translating some people has them regularly misidentifying instances of these two kinds—insisting (obviously mistakenly) that they have killed a grizzly bear, and warning of the rabbit protecting some winter kill—one would have reason to rethink one's translations. Such matters are perceptually obvious (and this is in large degree an empirical matter): people tend to get right such everyday matters about middle-sized physical and biological objects in plain sight, and with respect to which they have a significant practical interest in developing a competence. On the other hand, what is not perceptually obvious need not be treated as true under translation. The ill-glimpsed form in the brush might be misidentified, and this does not indicate a problem with translation. Similarly, translation that has us attributing glaring errors in reasoning of sorts that 'one would find obvious' should not be accepted, unless there are mitigating circumstances. (Factors that might count include the presence of drugs, alcohol, sleep deprivation, very strong personal interest in a conclusion other than that rationally indicated, and some kinds of defective training.) Empirical results regarding human foibles seem highly

significant when determining what perceptual matters should be relatively obvious in context.

Empirical results also seem significant for determining what forms of reasoning are cognitively obvious and which are not. People seem rather better at working with conjunction and negation than with conditionals. They seem better at working with conditionals when these involve concrete matters with which they have significant experience. They can believe the damnedest things when gods or governments are involved. They may believe contradictory things when that contradiction is 'well hidden'—so that it might require subtle proofs or particularly agile minds to appreciate. To insist that we 'preserve the obvious' in our interpretation is to insist that there are certain ways of reasoning that should be found in those we interpret:—the ones that characterize reasoning in the relevant set of critters (say humans). Some of these ways happen to be rational. This is a fact about human beings about which we are getting a progressively better grasp as we investigate human inferential tendencies. Since there is arguably significant human rationality, the advice to interpret so as to 'save the obvious' would have us interpret so as to find significant rationality (the obvious rationality). Still, to put it mildly, humans turn out to be subject to non-negligible irrationality. The rational principles that would serve as a corrective to such tendencies are, emphatically, not generally obvious to folk in context. So, if we must preserve the obvious, then findings of such irrationality (cases of forms of irrationality to which folk are given, cases where the contrasting form of rationality is not obvious), are no strike against an interpretation.

What is crucial on Quine's understanding of the principle of charity is that while there are real substantive constraints on interpretation, the substantive constraints here do not constitute a kind of a priori constraint. What is obvious perceptually or cognitively is an empirical matter: it is a matter of psychological tendencies. It is a matter regarding which we have significant empirical access rooted in everyday and common experience (of common perceptual capabilities and limits, and of common intellectual capabilities and foibles). It is a matter subject to systematic study (as in the empirical work on human inferential strategies and errors). On Quine's understanding, the charitable constraint that we preserve the obvious provides substantive constraints on interpretation only when conjoined with such empirical information—so that the substantive constraint here does not constitute an a priori constraint on interpretation to the effect that we must find rationality. Rather, if anything is a priori demanded here, it is that we should seek to find others reasoning in ways characteristic of the class of cognitive systems to which they belong (or characteristic of such critters in relevantly similar circumstances). The principle of charity is thus understood as an empirically informed constraint on interpretation, one that results from the application of our evolving empirical understanding of the relevant cognitive systems. 'The translator will depend early and late on psychological conjectures as to what the native is likely to believe' (Quine, 1987: 7). Since human beings are given to some significant forms of rationality (as well as to some significant forms of irrationality), the *derivative* demand that our interpretation be informed and conditioned by these expectations for some forms of rationality (and some irrationality) can be termed a principle of charity. We are to seek to find rationality of the common sorts—and failure to do so results in an account which is likely mistaken (for when the errors attributed to folk are highly unlikely, mistaken interpretation is relatively likely).

To be fair to the rationalists, one must notice that they would typically acknowledge a role for empirical information about human cognitive capacities and incapacities. Such information is acknowledged to be important in determining what interpretation is the best interpretation of an agent or people. Thus, Davidson (1984c, 1984d) insists that all attributions of error and irrationality count against an interpretation in

some measure (this is the a priori part), but that some count more strongly than others (this is at least partly an empirical matter). So, in determining what is the best interpretation of an agent or community of agents, we seek to attribute no irrationality; but, as some attribution of irrationality will be unavoidable, we should settle for attributions of irrationality that violate our learned expectations the least—i.e., the empirical information contributes to the negative weighting of attributions of error. (Of course, one would also want to allow that expectations will continue to evolve over time.)

This seems reasonable, but it also seems to make for a more attenuated form of rationalism. It originally seemed as if the strong rationalist could insist that there are certain levels and forms of rationality that would need to be found under interpretation. If this much could be taken to be a priori, and if the levels and forms of rationality corresponded at least to [L^r], then [L^r] becomes unassailable (at least at the level of competence)—and a RP is vindicated. But once one allows that empirical expectations can modulate the putatively a priori demand for finding rationality under interpretation, it becomes less clear what is *a priori guaranteed*. As noted earlier, the rationalist tends to adopt somewhat hedged formulations at this point. Thus, Davidson writes of it being a priori that beliefs, desires and actions are *preponderantly rational*. Here it seems that RP/[L^r] serves to characterize an a priori ideal to which all adequate interpretation must approximate, from which no acceptable interpretation can diverge *too much.* It is then acknowledged that what counts as 'too much' is at least partially an empirical matter. (If it were wholly an empirical matter, then again it seems that the a priori element here becomes vacuous.) So, as long as the 'too much' is not much, interpretations will need to conforms largely to [L^r], and background interpretations will not provide the basis for any but minor revisions of [L^r].

The contrast boils down to this: On the Davidsonian understanding (the strong rationalist understanding) the principle of charity articulates a powerful a priori constraint on interpretation, an invariant a priori ideal to which all interpretation must approximate. What makes for the best approximation may be empirically informed, as expectations for human capabilities and limitations may inform what errors make for significant divergence and which do not, but the model of reasoning to which the interpretations must ultimately approximate is invariant. On the Quinean understanding, that model of reasoning to which our interpretations must find our subjects approximating is neither a priori nor invariant. It is rather our evolving empirical understanding of human reasoning. As our understanding of human reasoning tendencies evolves under work in cognitive psychology (for example), the resulting expectations for both rationality *and irrationality* form a composite model of human cognition, and this model is that to which interpretation should conform for now. An element of this model—say, the expectation for certain forms of valid deductive reasoning, or the expectation for certain (fallacious) overuse of some judgment heuristic—is a piece of the model because it is a piece of our present best understanding of humans, not because it is given a priori as rational.[8]

To illustrate the difference between the strong rationalist approach and the empiricist approach, we can contrive a cartoon history of our interpretive practice. So suppose some point in that practice that should surely be congenial to the rationalist. Suppose that there were a time in which interpreters had no empirical expectations regarding human reasoning—only a normative model that includes things like normative decision theory, statistical reasoning, basic logic and the like. (I doubt that there ever was such a point, but let us not pause over this point.) According to the strong rationalist, this is not too impoverished a position from which to begin, for they insist that interpreters yet have the a priori ideal to which all interpretations must approximate anyway. Admittedly, interpreters would have no nuanced way of weighting divergence from the ideal—no empirical weighting of errors. But, we may

suppose that they might then count apparent divergences equally against an interpretation—and decide on an interpretive scheme for a people by choosing that scheme that minimizes divergence. But at this point there is a wrinkle to consider. On the one hand, interpreters could go on ever modifying their interpretations so as to 'explain away errors', adding ever more epicycles to their interpretive schemes, or they could have some sense of 'reasonable' or 'plausible' complexity. At some time, they may sense that an error in inference, an unacknowledged inconsistency or some other piece of irrationality is 'more likely' then yet another sense of the relevant terms in the subjects' lexicon, yet another epicycle. Perhaps, drawing on analogies with their own reasoning, they may sense that the avoidance of attributions of error is making for an unrealistically baroque interpretive scheme. It would be natural to think of such judgments as empirically informed, as drawing on courses of experience. But this is no problem for the strong rationalist who is ready to acknowledge that empirical information can help ascertain when an interpretation closely enough approximates to the a priori ideal. All that has been supposed here is that at some point interpreters do not feel a priori obliged to continue to complicate their interpretations to avoid yet another attribution of inconsistency.

So interpreters now find themselves attributing some irrationality to their subjects: the subjects are found to be marginally diverging from the ideal that was initially supposed. Significantly, they will find a pattern in the divergence; they will find that there seem to be systematic ways in which folk diverge from the normative model. Interpreters will also come to appreciate much about when folk in a given social context. learn certain things, for example, and what patterns of motivation are prevalent within a society or within a profession within that society (recall the EE that seemed relevant when thinking about Tversky and Kahneman's work). As suggested earlier, such empirical expectations serve to constrain an interpretation, and can add to the confidence of researchers that their interpretations are reasonable and that the divergences from some rational ideal that they seem to find are indeed real and systematic. As a result, investigators will come to have empirical theories or expectations having to do with human rationality and irrationality.

To this point, the cartoon history has been developed in a way that is highly favorable to the strong rationalist. For purposes of illustration I have supposed that the sort of normative model that the strong rationalist envisioned as anchoring interpretation does indeed constrain interpretation, at least at a mythical beginning in which no empirical expectation regarding human reasoning is brought to the table. Now we can let the disagreement between rationalists and empiricists emerge.

Suppose that we now undertake to understand some new agent or people. According to the rationalist, the model that is presupposed—from which attributed deviations are to be counted against the interpretations that we will entertain—continues to be the same, invariant, normative model (logic, statistical methods, the rest of normative epistemology, and decision theory). What has changed over time is the empirical background understanding which may influence how we weight the seriousness of attributed divergence from this model (but any divergence counts in some measure against the interpretation). So, when we find agents to be reasoning in defective ways that we have come to expect, this counts against our interpretation, at least a little. According to the strong rationalist, there is always some 'tax' on any attributions of irrationality—so that an interpretation that proceeds smoothly and corresponds to our empirically informed expectations for certain forms of irrationality will have yet thereby incurred an 'error tax' on its acceptability. Further attributions of irrationality—even if they conflict with no empirical expectations, and even if these expectations conform to EE-like expectations—may then be difficult to sustain. The a priori normative model continues to anchor and constrain our interpretations—and error taxes on attributions of irrationality preclude

interpretive findings that are too much in divergence from that model. Such is the strong rationalist picture.

In contrast, the empiricist need count little more than RRP as conceptually grounded. As expectations for ways of reasoning, including forms of irrationality and forms of rationality, emerge in the course of empirical work drawing on acceptable interpretive schemes, these constitute *an evolving model of human cognition—one that then serves for the empiricist as the model from which divergence is counted against an interpretation.* On this view, there is no 'tax' on attributions conforming to these expectations, this empirical model, even where these diverge from the normative model from which (for purposes of illustration we are supposing that) earlier interpretation took its departure.

Focus now on the issues left hanging at the close of our discussion of Kahneman and Tversky's challenge to RP/ [L^r]. The suggestion was that a range of empirically informed expectations of a diverse sort—EE— serve to provide support for an interpretation that has our subjects systematically violating [L^r]. It would seem that such results can accumulate so as to support an understanding of human cognition that is deeply at odds with [L^r]. The projected rationalist response was that such expectations were themselves dependent on interpretations and thus hostage to [L^r], so that deep challenges to [L^r] were foreclosed. In effect, while [L^r] might be given some 'tweaking', it remains an invariant a priori attractor to which all interpretive results remain tethered. The strength of the tether may vary somewhat with empirical results, but these themselves remain conditioned by interpretations tethered to [L^r].

The empiricists think differently of the fundamental constraints on interpretation. On their view there is no such a priori, invariant and substantive, model serving as an attractor for interpretation (nothing beyond something like RRP). As empirical understandings of human reasoning evolve, so does the 'attractor' to which interpretations must approximate—for those understandings constitute the model that informs interpretation as an evolving attractor. Because the background empirical expectations—EE—are not tethered to an invariant model along the lines of [L^r], there is no a priori guarantee that these background interpretations, and the interpretations/inquiries that they support, cannot give rise to deep challenges to [L^r], in fact they themselves could already reflect deep revisions of [L^r]. Thus, on the empiricist understanding, there can be adequate interpretations that attribute deep violations of [L^r], and that thus can occasion revision in [L^r] treated as an account of human cognition.

How can one decide between the two understandings of the principle of charity on offer—the strong rationalist understanding and the empiricist understanding (itself compatible with a weak a priori element such as RRP)? To settle the matter would require an extended reflection on the considerations adduced in a range of interpretive inquiries (such as those found in history and cultural anthropology) and on a range of investigations that suppose and sometimes reconsider interpretations (such as careful work in cognitive psychology). For my own part, I am convinced that the empiricist model provides the most adequate and best-motivated understanding of the relevant inquiries, but developing the support for this conclusion is beyond the scope of this article.

REFERENCES AND SELECT BIBLIOGRAPHY

Bealer, G. (1987) 'The philosophical limits of scientific essentialism', *Philosophical Perspectives* 1: 289–365.

Burge, T. (1979) 'Individualism and the mental', *Midwest Studies in Philosophy,* 4: 73–121.

—— (1992) 'Philosophy of mind and language: 1950–1990', *Philosophical Review,* 101: 3–51.

Chalmers, D. (1996) *The Conscious Mind*. Oxford: Oxford University Press.

—— (2002a) 'Sense and intension', in J. Tomberlin (ed.), *Philosophical Perspectives 16: Language and Mind*. Oxford: Blackwell, pp. 135–82.

——— 2002b. 'The Components of Content', in D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: Oxford University Press, pp. 608–33.

Davidson, D. (1980a) 'Actions, reasons, and causes', in *Essays on Actions and Events.* Oxford: Clarendon Press, pp. 1–19.

——— (1980b) 'Mental events', in *Essays on Actions and Events.* Oxford: Clarendon Press, pp. 207–25.

——— (1980c) 'Towards a unified theory of meaning and action', *Grazer Philosophical Studies,* 2: 1–12.

——— (1982) 'Paradoxes of irrationality', in R. Wollheim and J. Hopkins (eds), *Philosophical Essays on Freud.* Cambridge: Cambridge University Press, pp. 289–305.

——— (1984a) 'Radical interpretation', in *Inquiries into Truth and Interpretation.* Oxford: Clarendon Press, pp. 125–40.

——— (1984b) 'Belief and the basis of meaning', in *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press, pp. 141–54.

——— (1984c) 'Thought and talk', in *Inquiries into Truth and Interpretation.* Oxford: Clarendon Press, pp. 155–70.

——— (1984d) 'On the very idea of a conceptual scheme', in *Inquiries into Truth and Interpretation*. Oxford: Clarendon Press, pp. 185–98.

Henderson, D. (1987) 'The principle of charity and the problem of irrationality', *Synthese,* 73: 225–52.

——— (1990) 'An empirical basis for charity in translation', *Erkenntnis* 32: 83–103.

——— (1993) *Interpretation and Explanation in the Human Sciences*. Binghamton: State University of New York Press.

——— (1994) 'Conceptual schemes after Davidson', in Gerhard Preyer, Frank Siebelt and Alexander Ulfig (eds), *Language, Mind, and Epistemology: On Donald Davidson's Philosophy*. Dordrecht: Kluwer Academic Publishers, pp. 171–97.

Jackson, P. (1998) *From Metaphysics to Ethics: A Defense of Conceptual Analysis.* Oxford: Clarendon Press.

Kripke, S. (1972) *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Putnam, H. (1975a) 'The meaning of meaning,' in *Mind, Language, and Reality: Philosophical Papers,* vol. 2. Cambridge: Cambridge University Press, pp. 215–71.

——— (1975b) 'The analytic and the synthetic', in *Mind, Language and Reality*. Cambridge: Cambridge University Press, pp. 33–69.

Peacocke, C. (1992) *A Study of Concepts.* Cambridge, MA: MIT Press.

Risjord, M. (2000) *Woodcutters and Witchcraft.* Albany: State University of New York Press.

Quine, W. (1953) 'Two dogmas of empiricism', in *From a Logical Point of View*. Cambridge, MA: Harvard University Press.

——— (1960) *Word and Object*. Cambridge, MA: MIT Press.

——— (1970) 'Philosophical progress in language theory', *Metaphilosophy* 1: 2–19.

——— (1981) 'On the very Idea of a third dogma', *Theories and Things*. Cambridge, MA: Harvard University Press, pp. 38–42.

——— (1987) 'Indeterminacy of translation again', *Journal of Philosophy* 84: 5–10.

Rosenberg, A. (1985) 'Davidson's unintended attack on psychology', in E. LaPore and B. McLaughlin (eds), *Actions and Events: Perspectives of the Philosophy of Donald Davidson*. Worchester, MA: Blackwell, pp. 399–407.

——— (1988) *Philosophy of Social Science.* Boulder, CO: Westview Press.

Stich, S. (1990) *The Fragmentation of Reason*. Cambridge, MA: MIT Press.

Tversky, A. (1975) 'A critique of expected utility theory: Descriptive and normative considerations', *Erkenntnis,* 9: 163–73.

Tversky, A. and Kahneman, D. (1974) 'Judgments under uncertainty: Heuristics and biases', *Science* 185: 1124–31.

Whorf, B. (1956) 'The punctal and segmentative aspects of verbs in Hopi', in J.B. Carroll (ed.), *Language, Thought and Reality: Selected Writings of Benjamin Lee Whorf*. Cambridge, MA: The MIT Press, pp. 51–6.

NOTES

1 Putnam (1975a) provides a useful discussion of this tradition, which he then criticizes.

2 There is, of course, an obvious possibility which would take the air out of the neo-rationalist program: the element of the semantics of concepts that is accessible to one who is conceptually competent, merely by virtue of that person being conceptually competent, might turn out to be such a wimpy component of the

conceptual semantics that traditional empiricist deflationary responses seem appropriate.

3 Although the transition just now suggested reflects a grounds for skepticism regarding whether there is *any* line to be drawn between conceptual truths (analytic claims) and truths that are central to empirically supported theories of some matter (which would be synthetic). Famously, Quine (1953, 1960) argued that central elements of our empirical theories or understandings may seem relatively safe from revision, but that this matter of degree should not be confused with the supposed status of being 'true by meaning' or being 'purely conceptually grounded'—or any status that would make for a prioricity.

4 And to cut down on alternative understandings of choices/undertakings, we are to progressively constrain our understanding of these in terms of wider sets of standing beliefs and desires attributable to the agent or agents.

5 Those familiar with Rosenberg's work will doubtless find is strange to read of him as advancing a 'rationalist' position. This is a function of the early choice to treat the claim that there are significant a priori principles as a mark of rationalism, even when these principles are understood as conceptually grounded. As explained, many contemporary self-labeled 'rationalists' fall into this camp—for example, Peacocke, Chalmers and Bealer. Using the designation 'rationalist' is so broad a fashion that one can be an empiricist and yet still be a rationalist. Rosenberg would be a case in point. He thinks that there are significant, conceptually mandated, a priori constraints on interpretation—ones that are 'almost definitional' of action and related concepts—and then he insists that this renders such concepts unworkable for any respectable science. The conceptual constraints certainly do not seem trivial, as they amount to the idea that actions, beliefs, and desires interact so as to largely conform to the rather elaborate dictates of decision theory. Since he thinks that the social sciences are so constrained, he thinks that these are not respectable sciences, and would have us change the subject of inquiry. The same verdict is applied to any intentional psychology. In effect, Rosenberg accommodates the rationalist brief presented here by insisting that such concepts and constraints have no place in any respectable empirical science. We do not study unicorns—for good empirical reasons. Neither should we study actions and reasons.

6 Of course, the rationalist would insist that [L^r] serves as an a priori truth regarding all agents—all who act for reasons—not just humans. But, given that humans are supposedly such agents, it would need to serve as an a priori truth of human cognition: *if* humans have beliefs and desires, *if* they act, then [L^r] must (on the rationalist account) hold true of human cognition.

7 For further development of these themes, see Henderson (1991).

8 For a more sustained development of this contrast, see Henderson (1993), chapters 2–3.