1

Interpretation and Explanation: Claims and Concerns

This book is devoted to developing complementary accounts of both interpretation and explanation in the human sciences. This is undertaken with the conviction that an account of either interpretation or explanation will be inadequate and incomplete without a supporting account of the other. Accounts of either matter ultimately will need to appeal to, or suppose, important points regarding the other. On the one hand, explanatory success is itself the central desideratum of correctness in interpretation, and thus an account of explanation figures directly in articulating adequacy conditions for interpretation. On the other hand, explanations in the human sciences typically make reference to conditions described in intentional terms, and thus are dependent on interpretations. In view of this, philosophers have often attempted to repudiate accounts of explanation in the human sciences by advancing an account of interpretation that precludes certain things supposed in the account of explanation under attack. A satisfactory account of explanation must have the resources to deflect such criticisms.

In view of these connections, one could almost say that accounts of interpretation and explanation in the human sciences describe two faces of the same coin. However, writers who would make this claim have often been *methodological separatists*, holding that, as a result, the logic of explanation (and testing) in the human sciences is fundamentally different from what we employ in natural sciences. They have held that the constraints on interpretation are unique and quite disanalogous to the constraints on the description of phenomena of interest in the natural sciences. This is thought to foreclose the possibility of the sort of explanation appropriate to the natural sciences. For example, the possibility of such explanations is sometimes thought to be foreclosed by interpretive constraints rendering generalizations regarding intentional states

inherently non-nomic. Such a view is suggested by Davidson's (1980c, 1980d) doctrine of the anomalousness of the mental, the separatist implications of which have been pointed out by Rosenberg (1985b). Alternatively, it is sometimes argued that such explanations become superfluous in view of the intimate relation between intentional interpretation and a distinctly intentional form of explanation, (Winch 1958; McDowell 1985). What remains, it is said, is the possibility of a distinct sort of explanation, sometimes called empathetic understanding or rationalizing explanation, which is thought not to rely on nomic generalizations in the way characteristic of explanations in the natural sciences.

I, however, deny such separatist claims. I defend the *methodological naturalist* thesis that, at an important and fundamental level of analysis, the logic of inquiry (including explanation and testing) in the human sciences is the same as that informing the natural sciences. I do this, not by ignoring what is surely the strength of the separatist's position—the concern for interpretive understanding and a sensitivity to its central role in the human sciences—but by developing a methodological naturalist account of such interpretation. I show how debates over interpretations in the human sciences reflect a concern for the explicability of those interpreted, where "explicability" can best be understood in terms of a generic notion of explanation appropriate to the natural sciences as well as the human sciences. I proceed to develop the needed account of explanation in the human sciences, including an account of the much misunderstood status of rationalizing explanations in particular.

The point of the present chapter is to provide the reader with an overview of my position, which must be elaborated by a serial consideration of a set of interdependent issues: the nature of interpretation and the constraints on interpretation; the nature of mental states and whether it is a priori that, within any individual, they are preponderantly rational; the place of rationalizing explanations in the human sciences and the limitations of such explanations in these contexts; the nature of nomic generalizations and their place in explanation; the relations between the special sciences and the more fundamental sciences; the relation of psychological states to physical states; and so on. However, due to the intimate interrelations between such issues, and the interdependence of positions taken on such issues, dealing first with any one is bound to give rise to worries concerning the others, and to a resulting unease regarding the point then under discussion. My strategy here is simply to let the reader know what my general position on

the interrelated issues is, without defending that position at any length. As I then focus on particular issues in the chapters to follow, the reader can at least envision how I would seek to address threatening problems arising from the perspective of the other issues. Accordingly, this chapter is not intended to convince my readers, but only to inform them of what to expect. What I give here is a series of "I.O.U.s," which I discharge in later chapters.

In the next two chapters, I will provide my basic account of interpretation. This is a particularly important discussion, for it will undercut important separatist arguments. As noted above, it is commonly held that interpretation is alien to the general scientific method. There are constraints on interpretation as the attribution of intentional states; these are (more or less implicit) adequacy criteria reflected in interpretive practice. Such constraints have to do with what it is to "make sense" of behavior, what it is to find what is done "intelligible." It is these constraints that are often thought to give rise to a fundamental difference between the human sciences and the natural sciences, for they are thought to be essentially unlike constraints on the description of physical and biological systems. After discussing, in chapter 2, what has come to be the standard or common account of the constraints on interpretation, I will argue, in chapter 3, that a superior account of such constraints is possible, and that such an account does not have the threatened separatist implications.

There is, of course, some disagreement regarding the nature of these constraints, but two general points seem to enjoy a fair consensus. First, interpretation is a holistic matter—what interpretation is proper for a particular bit of text or behavior is dependent on what interpretation can be settled on for the larger text or range of the agent's behavior, of which the particular bit is a part. This, of course, is a familiar orthodoxy in hermeneutics; it is acknowledged in discussions of the "hermeneutical circle" in which the constructions placed on the parts and on the whole are "played off" each against the other in a continual process of refining interpretation. It is also orthodox in contemporary analytic philosophy of language and mind, where Quine and Davidson, among others, have described how the content of a belief is dependent on its place in a pattern of other beliefs and attitudes. It is also an integral feature of any philosophy of psychology and mind that is influenced by functionalist accounts of cognitive processes. I am convinced both that this first point is correct and that it need not give rise to separatist results. Holism, after all, is not unique to interpretation.

In fact, it is characteristic of the analysis of systems. For example, it is found in the analysis of biological systems, and functionally characterized systems generally, as well as in the treatment of intentional or cognitive systems.

Second, it is commonly suggested that in "finding intelligible" and "making sense" of a range of behavior, the holistic pattern employed is, of necessity, a *fundamentally charitable* one. This is to claim that interpreters are under a *basic methodological constraint* to interpret so as to construe people as predominantly rational in thought and deed, and as largely correct in their beliefs. Such a "principle of charity" is, I believe, best understood as a proffered codification of interpretive practice. Again, I think there is wide agreement on the claim that interpretation is constrained by such a principle. Indeed, it is easy enough to understand "making intelligible" and "making sense" to mean "finding the rationality in" the behavior at issue. This fits comfortably with the view that interpretation automatically leads us to rationalizing explanations of the thought and deeds of those interpreted. Such views find expression in the writings of Davidson, Winch, Turner, and Taylor, among others.

This second point of substantial consensus regarding the constraints on interpretation is particularly significant. For it is what gives rise to the view that interpretation is really fundamentally distinct from all other sorts of inquiry. Were the principle of charity really a fundamental methodological constraint on adequate interpretation, this would make for a deep difference between interpretation and the description of phenomena in the other sciences. After all, it is said, such a constraint "has no echo" in other contexts (Davidson 1980d, p. 231). Sure, the scientific description of phenomena is "theory laden" in all contexts, and is thus constrained by extant theoretical understandings. But this theoretical background is supposedly responsive to standard scientific concerns, for explanatory power, for example. As our theory changes, our descriptions of phenomena come to be informed by different theoretical principles, and thus different constraints, without thereby putting us in violation of any fundamental methodological constraint. In contrast, interpretive description is commonly thought to be constrained by the principle of charity in a more fundamental way. These constraints are thought not to be just a matter of interpretation being theory-laden description, for this would make such constraints subject to the shifting sands of theory. Instead, the constraints are more a matter of interpretation being wedded to an a priori restricted range of

psychological theories, those that view people as preponderantly rational (Root 1986). Or perhaps it is preferable to say that the constraints on interpretation themselves force that range of theory on us. On this view, one might say that it is not that our descriptions must lead to explanations, but that they must lead to (mostly) *rationalizing* explanations.

I have no quarrel with the principle of charity when it is understood as a crude codification of interpretive practice and thus subject to substantial refinement. For it certainly captures important aspects of that practice. However, I argue that it is a *derivative principle*, a general rule of thumb, the place for, and limitations of, which can be understood in terms of a yet more fundamental constraint that I have (1987a) called the principle of explicability. The principle of explicability councils us to so interpret as to construe people as doing explicable things, in view of their beliefs and desires, and as believing and desiring explicable things in view of their other beliefs and desires and their training.

There are a range of considerations supporting this view of charity in interpretation as derivative, and thus not a fundamental methodological constraint. To begin with, when one considers refinements philosophers have come to make in formulating versions of the principle of charity, one finds that these ultimately are tailored to what is explicable in the way of cognitive and practical successes and failures, not to degrees of normative propriety. Those cases in which the principle of charity seems most constraining are just those where error would be generally most inexplicable, and attributions of error would most tend to violate the principle of explicability. Further, when one attends to anthropological controversies concerning what is an ultimately acceptable interpretation in cases of apparent irrationality or egregious error, one finds that the issue quickly comes to be the explicability of what is said and done according to the competing interpretations, not the normative propriety of what is attributed in the competing accounts. Such observations (to be defended later) show that the principle of charity is tailored to fit the dictates of the principle of explicability, given present general descriptive information about psychological and sociological processes. This indicates the derivative character of the principle of charity. When properly qualified, it can thus be recognized as just a reflection of the theory-laden nature of description, given present theoretical resources. (Of course, the notion of explanation supposed here is the standard one: roughly, explicability in terms of causal antecedents, where these are picked

out in terms of background nomic generalizations—here regarding human cognitive capacities, learning and socialization, and so forth.)

I believe that the sort of argument suggested just now (and developed in chapters 2 and 3) demonstrates that the adequacy of an interpretation turns on the explicability of what is attributed to those interpreted. Obviously, this is an encouraging result for a methodological naturalist. For, if the methodology of interpretation is compliant to the concerns for explanation, and the relevant sort of explanation is a generic sort that is also instanced in the natural sciences, then what once appeared to be deep methodological differences turn out to be just central bits of psychological theory reflected in theory-laden descriptions. However, a survey of the literature will show that there remain two related sources of misgivings, both of which are variants of the general claim that while interpretation may follow explicability in many cases, it is still subject to a special constraint to attribute rationality (and perhaps correctness) at some minimal level.

First, an a priori minimal rationality requirement on interpretation is sometimes defended by arguing that it is "constitutive of the concept" of intentional states, and thus an adequacy criterion for the attribution of such states, that those states be minimally rational (Davidson 1980c, 1980d, 1980e; Root 1986; Stich 1985). Here the suggestion is that, explicability aside, if our attributions do not uncover enough rationality, then whatever we are identifying simply are not intentional states. (If we seem to be attributing less than the minimal degree of rationality, we have, perhaps covertly, become eliminativists of some stripe.) As a result, strictly speaking, we are then not interpreting and we are not addressing the subject matter of the social sciences and intentional psychology. The idea, I suppose, is that the principle of charity can bend somewhat with empirical findings concerning limited irrationality, but ultimately, it sets certain fixed limits on the attribution of intentional states. Another way of expressing these concerns is to insist that charity limits what empirical findings there could be regarding irrationality, and thus it limits what generalizations we might have to explain irrationality in thought and deed. Such limits are thought not to be empirically fluid. So concerns for explicability may tailor charity, but ultimately only minor alterations are possible.

In the fourth chapter, I rebut such arguments. First, I show that they overestimate what appeals to constitutive criteria can accomplish. Ultimately, if philosophically defensible, such appeals have to do with the centrality of certain principles to our present

theoretical resources. As a result, they do not give rise to an absolute prohibition on the development of concepts in particular ways. Minimal rationality requirements cannot be more constraining on future developments in psychology than fundamental Newtonian principles ($e = 1/2mv^2$, for example) were constraining on the subsequent development of physics. Second, it is significant that contemporary psychological resources may themselves fail to support the claim that beliefs and desires must be *predominantly* rational, depending, of course, on how rationality is itself judged. Finally, the stubbornness of the constitutive criteria supposed to give rise to the rigid minimal rationality requirement is typically understood as the result of a variant of the principle of charity that is perhaps informed to an extent by the principle of explicability, but that is not fully subservient to explicability. Here the considerations adduced in chapters 2 and 3 may be employed to show that this is untenable, for there is simply no good reason for thinking that charity and explicability are ever competing desiderata for social scientific accounts, as they would be if the principle of charity were not freely tailored to explicability.

The second defense of minimal rationality requirements is to insist that the explicability of concern in interpretive contexts is predominantly a matter of explicability in terms of rationality. This is to say that interpretive work must lead us to predominantly rationalizing explanations of those we interpret, perhaps because such is the sort of explanation associated with intentionality. Ultimately, when pressed, such a general position is probably not distinct from that discussed in chapter 4, except as a matter of emphasis. However, in discussions of social scientific explanation in particular, such a position is occasionally put forward with relatively little discussion of charitable constraints on interpretation. In some cases, it is advanced not as a concomitant of intentional interpretation and explanation as such, but as the proper sort of explanation for the social sciences (as distinct from psychology). In this form, it finds an advocate in Jarvie (1964), who insists that such explanations are fundamentally a matter of uncovering "the logic of the situation," and who insists that such a focus is independent of psychological results. Turner's (1980) discussion of sociological explanation as a matter of giving account of the logic of "practices" may also be understood as supportive. The general claim that rationalizing explanation is a preferred form of intentional explanation, and must constitute the clear majority of such explanations, is discussed in chapter 5, where it affords the occasion for making

important points about the status of rationalizing explanation.

Rationalizing explanation has proven to be quite limited in some contexts—say the anthropology of religion—leaving much in need of further sorts of explanation. But, while this observation casts doubt on the adequacy of rationalizing explanation as the basis for the human sciences, it is really only a preliminary. The claim that rationalizing explanation must be supplemented (even heavily supplemented) does little to rebut the assumption that rationalizing is a particularly appropriate, fundamental, or preferred sort of intentional explanation. But an adequate account of the nature of rationalizing explanation and psychological explanation generally will lead us to repudiate the privileged status often accorded to rationalizing explanations. One particularly significant complementary alternative sort of explanation is the sort that appeals to cognitive strategies such as are presently the subject of much work in cognitive psychology. These strategies are commonly not optimal, and are sometimes systematically normatively inappropriate. When we explain various beliefs or actions by reference to such strategies, we are clearly employing a generalization-based explanation. We might call such explanations irrationalizing explanations. I draw on recent work in the philosophy of mind and psychology to argue that rationalizing explanation and irrationalizing explanations are on a par. Both implicitly or explicitly posit general cognitive capacities or liabilities characterized in terms of rules of reasoning. The only significant difference between the two is whether or not the rules used to characterize the relevant cognitive dispositions happen to formulate normatively appropriate ways of reasoning.

Together, chapters 5 and 6 present an account of rationalizing explanation that situates such explanations within a general account of the explanation of events. I argue that scientific explanations come in two complementary forms: answers to how-questions and answers to why-questions. Functional analyses provide answers to how-questions by accounting for sophisticated or complex capabilities (to maintain homeostasis or to solve a problem, for example) in terms of simpler dispositions (Cummins 1983; Rosenberg 1985a). Dispositions are characterized in terms of particular outputs being keyed to certain inputs. Thus, successful analyses in terms of dispositions uncover transition laws, nomic generalizations regarding a sort of system, which can be used in answering why-questions. Scientific explanations for events, singular causal explanations, are answers to why-questions. They identify causes of an event by

picking out what it was about the course of antecedent events that, had things been different in that respect, the explanandum event would (probably) not have obtained. Thus, scientific explanations of events draw on nomic generalizations that allow us to appreciate the causally relevant factors in the course of events. Such a general account of scientific explanation owes much to Salmon (1984) and Humphreys (1989a, 1989b). Now, both rationalizing and irrationalizing explanations, as described in chapter 5, will readily be recognized as cases where a more or less explicit functional analysis has given rise to transition laws regarding human cognitive dispositions that are supportive of causal explanations answering why-questions.

However, some will find that they have a stubborn suspicion that explanations in the human sciences proceed in terms of normative principles, not descriptive generalizations. McDowell (1985) articulates such a view. I argue that this cannot be. Nomic principles *qua* normative principles are irrelevant to answering why-questions, understood as accounting for the occurrence of an event of a particular type. Of course, when statements of such principles are used as representations of basic or acquired cognitive dispositions, they are quite relevant. But this is to transform the principle into a descriptive claim. It is then in this capacity as a descriptive generalization that the principle comes to support answers to why-questions.

My account of both interpretation and explanation in the human sciences requires that there be nomic psychological generalizations—psychological laws. Of course, whether there are such laws has itself been the subject of debate. Recent work in analytic philosophy has led some to the view that there are no nomic generalizations of intentional psychology. This claim has been supported in several ways. Responding to such challenges is the task of chapters 7 and 8.

Alexander Rosenberg has argued that at least certain crucial generalizations of intentional psychology cannot be nomic because they are not empirically refinable, and being so refinable is taken to be characteristic of nomic generalizations. According to Rosenberg, the rudimentary generalizations informing rationalizing explanation are involved in interpretation in a way that precludes their being empirically refined. To be so refinable, it must be in principle possible to obtain cases where the generalizations fail.[1] But, Rosenberg argues, apparent violations of such generalizations must themselves be treated as dubious, due to the

way in which these generalizations inform interpretation. Since prima facie counterinstances to the central principles must ultimately be taken as cases of poor interpretation and thus as spurious counterinstances, we could never have empirical grounds for modifying these central principles.

Rosenberg's challenge is taken up in chapter 7. The central difficulty with this argument is that it ignores the range of generalizations that inform interpretation, and, as a result, it ignores the ways in which other portions of this store of generalizations can be used to support interpretations that indicate the need for refinements on those rudimentary generalizations that Rosenberg believes are insulated from test. I develop this point by elaborating an account of bootstrap testing in the human sciences. (My debt to Glymour's (1980) account of bootstrapping will be obvious.) In bootstrap testing, a hypothesis taken from a theory may be tested by using other portions of the same theory to derive instances (or counterinstances) of that hypothesis from what is observed. Confirmation thus is recognized as a triadic relation: evidence confirms a hypothesis with respect to a theory. I am able to show how my account of interpretation allows us to appreciate the range of theory relevant to common psychological experiments and, thus, to appreciate how such experiments can be used in refining even those cherished basic principles informing rationalizing explanation. I illustrate the process by discussing concrete experimental work by Tversky and Kahneman, among others.

Davidson's (1980c, 1980d) discussions of "heteronomic generalizations" and his associated doctrine of the "anomalousness of the mental" pose a second fundamental challenge to the nomic status of generalizations in intentional psychology. Davidson's distinction between heteronomic and homonomic generalizations has to do with what is involved in refining generalizations. We may suppose that rough generalizations are expressed using only vocabulary from a particular level or type of description—physical, biological, or intentional, for example. Generalizations are heteronomic when they cannot be refined into strictly universal, non-ceteris-paribus, generalizations by employing vocabulary of the same basic sort employed in the rough generalizations.[2] Homonomic generalizations can be so refined. Davidson seems to suggest that heteronomic generalizations are all non-nomic; he compares them to Goodman's clearly non-nomic generalization concerning grue emeralds. When one sorts through Davidson's various discussions, one finds that his reservations ultimately suggest that

the features mentioned in heteronomic generalizations cannot be real causal factors (even though the events they characterize, including psycholological events, can be causal events). At most, then, heteronomic generalizations reflect there being a range of causal regularities underlying the rough regularity that they describe. According to Davidson, the mental qua mental can only be subject to heteronomic generalizations. It would follow that there can be no nomic psychological generalizations.

Related concerns have been expressed by Malcolm (1968) and Kim (1989a, 1989b). Kim formulates a "principle of explanatory exclusion" holding that "there can be at most one complete and independent explanation" for any given explanandum. With respect to psychology, the issue then becomes: how can a given bit of behavior be explained both by certain neurophysiological antecedents that we all admit cause it and by certain psychological features of the agent as well? If an adequate answer cannot be provided, then it looks as though we may be forced to conclude that psychological features cannot be causally relevant, and that psychological generalizations are, as a result, non-nomic. Kim's own reflections (1989b) lead him to conclude that, if a higher-level feature is to be causally relevant, it must be reducible to causally relevant lower-level features. Such reducibility is thought to render the two explanations not independent, thus bringing them into conformity with the exclusion principle. On this view, if psychological states are to be causally relevant, intentional psychology must be reducible to neurophysiology. Since such a reduction is widely acknowledged to be impossible, psychological states would seem to be causally irrelevant, as was suggested by Davidson.

In chapter 8, I take up Davidson's and Kim's misgivings, arguing that the heteronomic generalizations characteristic of the special sciences can be nomic and deal with causally relevant features, despite the fact that theories in the special sciences are, in an important sense, irreducible to lower-level theories. Thus, generalizations in intentional psychology can be nomic despite the fact they are not reducible to lower-level theories such as neurophysiology. To begin with, I build on Kim's own notion of a "supervenient causal relation." In such relations, certain higher-level features are causally relevant by virtue of "supervening on" and being "realized in" causally relevant features at a lower-level. I argue that this account of the causal relevance of higher-level properties is essentially correct. But, the demand that higher-level theories deal with supervenient causes does not give rise to a significant reductionism. I argue that

Kim's reductionism is either trivial or unacceptable, depending on how reduction is understood.

When the range of points broached in this introductory chapter have been developed and defended, we will have a full and well-integrated account of the human sciences focusing on the pivotal notions of interpretation and explanation. It is my hope that it will then serve as a defense of aspirations some of us hold for a *science* of human psychology and social life, and a defense that honestly answers some of the reservations responsible thinkers have had regarding such a science. In particular, I seek to provide such a defense by acknowledging the importance of interpretive understanding in the human sciences, and by developing an account of such understanding that both reflects interpretive practice and complements an account of a science of human beings. The marriage of interpretation and scientific methodology should be a happy one, I predict, for the parties share the same interest—explanation—and both need the other to pursue this interest.