ADRIAN M. S. PIPER

TWO CONCEPTIONS OF THE SELF

(Received 5 October, 1984)

The Humean conception of the self prevalent in the contemporary literature in moral and political philosophy, philosophy of mind, and action theory has yielded a persuasive model of human action that has contributed considerably to our understanding of moral motivation, rational action, and many other issues. But it has also generated certain problems. I should like to take issue with this conception, first by describing it in some detail and charting its connection with two such interrelated problems in moral psychology. Than I shall propose an alternative conception, cribbed in its essentials from Kant's metaphysics, that purports to do an even better job of explaining the psychological phenomena. Finally I shall argue that on the suggested alternative, these two problems do not arise.

I. THE HUMEAN CONCEPTION OF THE SELF

The familiar Humean conception of the self is structured and motivated by desire.[1] By a *desire*, I shall mean, provisionally, something like what Brandt and Kim seem to mean by a *want*[2]: i.e., a disposition to feel pleasure or satisfaction in thinking about or admiring the object of desire, and a disposition to feel disappointment or frustration in its nonattainment. On this conception, the self is to be identified with its most central desires, plans and projects – i.e., with what Bernard Williams calls its *character*.[3] These desires structure the Humean self in two ways. First, through the distinction into first- and second-order desires,[4] they determine our evaluation of the other elements of personality: our emotions, beliefs, and so on. According to this view, first-order desires are desires for particular states of affairs conceived as external to the self: for nuclear disarmament, for example, or for a piece of carrot cake. Second-order desires are desires for certain first-order desires, hence for their attendant thoughts, feelings and dispositions. Second-order desires are desires that one be (or become) a certain kind of person: they

Philosophical Studies **48** (1985) 173–197. 0031–8116/85.10

constitute a *desired self-conception*. For example, suppose I have a central first-order desire for sex, drugs, and rock and roll. This desire may fulfill a second-order desire to be the kind of person who desires such things. Or it may frustrate a second-order desire to be the kind of person who pines only after beauty, truth and goodness. The actual first-order desires which constitute the self either buttress or undermine our desired self-conception; our second-order desires tell us what that desired self-conception actually is.

Thus there is an important distinction to be drawn between a self-conception and a conception of the self. A *self-conception* picks out the basic intentional features in terms of which I actively identify myself. A *conception of the self*, on the other hand, provides a theoretical model that purports to explicate matters of fact regarding the nature and dynamics of the self. That I view myself as tactless is part of my self-conception; that I am in fact to be identified with my moral convictions or social relations or desires is part of a conception of the self with which I may or may not be in agreement. Thus the two are independent.

On this view, the Humean self is structured by its desires in a second way as well. For the importance of rationality as a defining feature of the self consists in its ability to provide hierarchical order and consistency to the totality of desires one has on any particular occasion: to ensure their mutual consistency with one another, to rank them in order of importance, to schedule a plan for their satisfaction with respect to value, probability, spatial and temporal proximity, duration, and comprehensiveness, and finally to facilitate their satisfaction through maximally efficient action.[5] The structural components of the self are desires, and the rational self is one in which these desires are ordered according to the canons of instrumental reason. Understanding and reason are thus subordinate means for satisfying our desires.[6]

Desires, on this view, structure not only the self but the actions in which it finds expression. It is claimed that we begin with a certain set of desires, and formulate beliefs about the most efficient means at our disposal for satisfying them. Other things equal, the actions we choose to perform then reflect those beliefs. Perceived or imagined objects of desire, then, provide the conative origin of all actions.[7]

This conception of the self can be described as *future-oriented* in the sense that the self finds expression and continuity in setting for itself, in the present, some future, desired state of affairs that it can anticipate working to

actualize over time.[8] This feature of the Humean self can be regarded as the consequence of tying a dispositional analysis of traits of character to the foundational notion of a desire.[9] To call a person generous or corrupt, on this analysis, is to describe a way she is disposed to act under certain circumstances. But on the Humean conception of the self, all action is motivated by desires the agent wishes to satisfy. Hence the concepts we invoke to describe a person's character or personality denote certain kinds of desires that person is disposed to try to satisfy under the relevant circumstances. The self then achieves full realization to the extent that it succeeds in satisfying those desires.

The Humean self is also *heteronymous*, to use Kant's term,[10] in that the conditions of its expression are objects or states of affairs perceived as temporally and/or spatially external to the self in its present incarnation. This external relation of the self to its desired objects generates actions performed for the sake of those objects.[11] And the full realization of the self consists in bringing into existence those extrinsic desired states of affairs.

Finally, the Humean conception of the self is *individualistic* in that as a Humean self I am motivated to satisfy some desire only if the desire in question is mine. If the desire belongs to someone else, than I am motivated to satisfy it only if I have a further desire I might thereby satisfy: i.e., to satisfy his desire. Of course this is not to say that all the desires I am moved to satisfy are inherently egoistic.[12] I am moved to satisfy my desire to advance the common good, even at considerable personal disadvantage, by the prospect of advancing the common good, not by that of personal satisfaction. Nevertheless, advancing the common good must be the object of my desire; otherwise I have no motivation for advancing it. Thus on this conception of the self, that I merely believe some state of affairs to best contribute to the common good, or to satisfy someone else's desire, is not sufficient to motivate me to try to achieve it. In addition, I must have a desire to so contribute; to do so must be the object of *my* desire.

These observations indicate the intimacy of the relation between the self and agency. If I were nothing more than a passive contemplator, I could have no self whatsoever. For if I necessarily failed to distinguish, among the ongoing panorama of events, some which I caused to occur, I would equally lack the means of identifying those among my experiences which were caused by something else; I could identify no subject to whom these events were hap-

pening. But if I were unable to distinguish myself from the events that happened to me, it is difficult to imagine how I might then distinguish my *self* at all.

However, that the self must find definition and expression through action does not imply that the self must be future-oriented, heteronymous, and individualistic. Hence it does not follow from the intrinsic connection between selfhood and agency that the Humean conception of the self is necessarily the correct one.

II. SELF-EVALUATION AND MORAL PARALYSIS

Next consider two related issues subsumable most broadly under the rubric of moral psychology. The Humean conception of the self generates a difficulty about the possibility of self-evaluation, as both proponents and opponents of that conception have recognized.[13] Essentially, the difficulty lies in the notions that the self is structured by first- and second-order desires, and that second-order desires provide criteria for evaluation of the motivationally effective desires of the self. The question immediately arises of why we should accept as authoritative criteria these second-order desires. Why should we not subject them, in turn, to the critical scrutiny of third-order desires, and so on, *ad infinitum*? Frankfurt's answer is that "It is possible ... to terminate such a series of acts without cutting it off arbitrarily", by identifying oneself *decisively* with one of one's first-order desires. This means that questions regarding higher-order desires do not arise:

> The decisiveness of the commitment [one] has made means that [one] has decided that no further question about [one's] second-order volition, at any higher order, remains to be asked.[14]

But *on what ground* has an agent made this decision? If there are no further grounds for halting the ascent to higher-order desires, then the decisive commitment one has made would seem to be arbitrary after all. That I lack the stamina or interest necessary for performing higher-order acts of self-evaluation does not confer authority on the $n+1$-order desires beyond which I refuse to look, any more than my refusal or inability to consider your point of view settles authoritatively the question of who has prevailed in our disagreement. If an authoritative termination of the infinite regress of orders of desire is to be contrasted with an arbitrary one, we shall need a better reason

for doing so than that we are too tired, or unwilling, to press further the hard task of self-evaluation.

Hence if the Humean conception of the self is the correct one, we should experience some difficulties in performing that task. For any set of desires and interests to which I commit myself is likely to seem arbitrary upon reflection. No action can then fully express my self because none can satisfy the desires of my self. And none can satisfy the desires of my self because there are no *n*-order desires with which I can fully identify. The consequence is a desired self-conception attenuated by doubts about the worth of that desire, and so about the action it is assumed to motivate.

This calls into question the extent to which a Humean self might be motivated to action at all. If the infinite regress inhibits one's rational self-identification with any *n*-order set of desires, then there can be no actions to which one can commit oneself wholeheartedly and without reservation – not necessarily because one has conflicting impulses, but rather because the worth of any such impulse is automatically subject to doubt. That I am not in fact left with a continuing case of moral paralysis that vitiates my capacity for decisive and principled action suggests that the Humean conception does not render accurately the psychological facts.

Some proponents of the Humean self seem to embrace a kind of moral paralysis as a sign of authenticity. Charles Taylor, for example,[15] seems to believe that it is both irresponsible and self-deceptive to presume that one's chosen action might successfully and conclusively quell the stirrings of conscience. Hence he accepts without reservation the implication that dogged and continuing reevaluation of the choices made by the self, and the principled doubt that any such reevaluation is itself adequate, must be permanent features of an authentic self. But I find troubling the notion that there are, and should be, *in theory* no terminating criteria for evaluating the worth of any desire one might have, nor of any action one might undertake. For then the whole point of ascending to the self-reflective stance of second-order desires in the first place seems to have been lost.

Others may feel no qualms about simply digging in their heels and coupling a forceful assertion of their intrinsic desires with a bald refusal to give any further justification of those desires. But this clearly fails to address the question of whether or not such terminating criteria have been met. We are ready to accept such a stance only when they have, in point of fact, been

met: The familiar intrinsic desires for friendship and intellectual stimulation resist further regress, whereas the anomalous or capricious desires to spend one's evening howling at the moon, or for continuing self-obliteration invite one. The diversity of our responses to such cases may, of course, be purely fortuitous. But it is more likely that the former set of desires is intelligible and the latter is not, and that both are susceptible to terminating criteria of rational intelligibility that the former satisfies and the latter violates.

However, to explicate these criteria and their relation to the lower-order desires they evaluate requires us to move beyond the scope of the Humean conception of the self. For by definition, the concept of a higher-order desire is insufficient to supply such an explanation; and this is all the Humean conception of the self has to offer. Thus suppose that there are rational grounds on which decisive identification with one's n-order desires are made. This insures the authority of the decision to terminate the regress at some particular point in the series, but only by sacrificing the evaluative authority of second-order desires. For whatever the ground on which we justify our decisive commitment to some set of n-order desires, those grounds cannot themselves be desires of any order. If they were, the regress could be reopened, merely by asking for reasons why we should be impressed with the authority of those $n + 1$-order desires. Here it will not do simply to point out that these are the desires we happen to have, or even that these are the final or intrinsic desires which confer urgency on all those that are instrumental to their satisfaction. For that we have desires doesn't demonstrate that they are nonarbitrary from the perspective of rational justification (suppose, for example, that my deepest intrinsic desire just *is* to spend my evenings howling at the moon). A *fortiori*, it doesn't demonstrate that they constitute authoritative and nonarbitrary terminating criteria of self-evaluation. Hence any such criteria to which we may appeal successfully must be independent, not only of the desires we actually do have, but even of those we should have. For part of the function of such criteria will be to furnish conclusive and compelling reasons for why we should have precisely those desires rather than some others.

Gary Watson[16] has proposed a conception of the self that addresses this requirement. He suggests that we distinguish Reason and Appetite as two independent sources of motivation, as Plato did. On Watson's view, Reason is the source of evaluative judgments about "those principles and ends which [one] – in a cool and non-self-deceptive moment – articulates as definitive of the good, fulfilling, and defensible life".[17] These constitute rational values

which are motivationally effective and from the standpoint of which the worth of our motivationally effective desires can be assessed. Since rational evaluations are of the first order too, the infinite regress does not arise.

Or does it? Watson's picture of rational values suggests that the regress is to be blocked by demonstrating that the ends "definitive of the good, fulfilling, and defensible life" are authoritatively justified, i.e. that it would be absurd or irrelevant to raise any further doubts about the rational value of those criteria. This much seems to follow by definition of "defensible". But this characterization thereby begs the question. For we can agree that the rational defensibility of certain final ends renders them immune to the pressure to push the regress of justification one step further. But merely *calling* them defensible does not *make* them defensible. Without knowing what Watson intends by "good", and to whom and under what conditions a life is "defensible", there is no reason why my most favored activity of howling at the moon should not be definitive of the "good, fulfilling, and defensible life" for me. And however ready you may be to accept my chosen way of life, surely you are justified in raising further doubts about its rationality. If Watson's rational values are truly rational, then we should be able to give persuasive reasons for holding them, and for according them precedence over the promptings of desire. That is, we should have some reason to believe that we are capable of evaluating ourselves *correctly*. Otherwise, Watson succeeds only in shifting the infinite regress from appetitive desires to "rational" values, rather than terminating it.

Watson not only does not furnish such criteria. In fact, he cannot. For in painting a bipartite conception of a self that includes two independent sources of motivation, he leaves open the psychological question of which source is in fact authoritative for any particular self, and begs the philosophical question of which source should be. He is concerned to emphasize that the Reason-Appetite distinction does not commit us to any necessary or inevitable split between reason and desire, since, for example, we may value certain activities, such as eating or sex, precisely because of the desires they satisfy.

> But the distinction does commit us to the possibility of such a split. If there are sources of motivation independent of the agent's values, then it is possible that sometimes he is motivated to do things he does not deem worth doing.[18]

However, even this understates the case. For if there are two, mutually independent *sources* of motivation within the self, then surely it must be an

open question with which source the agent identifies on any particular occasion, hence which constitutes her self-conception or (in Watson's terminology) "standpoint".[19] Watson seems to take it for granted that an agent must be identified with the values that come from reason, and dissociate himself from any desires or actions that do not conform to them. But this assumption underestimates the role of action as expressive of the self. When I perform genuine action, there is a state of affairs which I envision as its outcome, intend to bring about, and work to bring about. The "I" in the preceding sentence is not neutral between reason and desire. Whichever source of motivation is causing the action is the one that, for that moment, expresses my self. If desire is motivating the action, and reason disapproves of it, then so much the worse, for the time being at least, for reason. And if the conflict persists over the long term, so much the worse for the unity of the self.

Hence the problem of moral paralysis resurfaces in the form of a dilemma for the Platonic bipartite self: which part of the self ought to have motivational priority on any particular occasion? And who – or what – ought to settle this question? If I act on my desires at the expense of reason, reason can reproach me with incontinence; or, at worst, Aristotelian self-indulgence. If my rational values take motivational precedence over my desires, the approval of conscience may be insignificant in the face of the frustration, regret, and alienation contingent upon ignoring the acknowledged demands of desire.[20] If I am unlucky enough to be torn by equally strong but conflicting tendencies from reason and desire, I may be as fully paralyzed as Buridan's Ass, and for much the same reason. If not, I will in any case be unable to exercise my agency in determining my behavior, and so will suffer the disquieting experience of being propelled into action by forces external to my will, *regardless of the course of action on which I finally embark*.[21] Under such conditions of perpetual internecine conflict, it is a wonder that we manage to do anything at all.

And so for Watson's Platonic conception of the self, the psychological problem of moral paralysis is not resolved but exacerbated. This conception fails to resolve the problem because it contains an unexplicated assumption about which feature of the self has authoritative and motivational priority. Hence his proposed solution to the problem of self-evaluation suffers accordingly. If reason and desire must vie for control of the self as the original picture seems to suggest, then to appeal to rational values to terminate the

proliferation of orders of desire is no less arbitrary than it would be to appeal to any appetitive desire to do so. But in the absence of any further, highest court of appeals within which these conflicting demands can be adjudicated, a rationally and morally imperfect agent who nevertheless acts decisively and well much of the time must remain a theoretical enigma.

Thus if we cannot provide, even in theory, some such terminating criteria for self-evaluation, it is unclear why we should bother to evaluate ourselves in the first place. Without an authoritative justification of the values and norms on which we both act and rely for criteria of self-evaluation, there is no non-arbitrary reason why we should commit ourselves to those values rather than to some others. Then it is not easy to explain how or why our actions and character *should* matter, either to us or to anyone else, at all.

III. A KANTIAN CONCEPTION OF THE SELF

Clearly, the problems of self-evaluation and moral paralysis can be generated by any multipartite conception of the self. Just as clearly, these problems are also generated by a unipartite conception of the self as structured and motivated by desires alone. The question then arises of whether the remaining in-house candidate, namely reason, might be adequate to structure and motivate a unipartite conception of the self that both successfully circumvents these problems and respects the psychological data.

I shall argue that reason fulfills these desiderata, first by limning what I shall describe as a *Kantian* conception of the self. On this conception, roughly, the self is motivated and structured by the internalized norms that dispose it to various kinds of conscious behavior; and overridingly by a highest-order norm of theoretical rationality that secures its internal unity. My debt to portions of the Analytic and Dialectic of Kant's *Critique of Pure Reason* will become increasingly evident,[22] as will my frequent departures from Kant's actual doctrine.[23] Later, in Sections IV and V, I shall try to show the competitive superiority of the Kantian conception, by arguing that it better explains certain psychological facts of our experience, and also provides solutions to the companion problems of self-evaluation and moral paralysis.

Consider first the question of motivation. Kantians often confront the objection that without stipulating desire as a motivation for action, they are hard pressed to provide an explanation of why an agent would act in accordance with norms or principles. Elsewhere[24] I have argued that desire in any

case cannot be a necessary motivation for action if the concept of desire is nontrivially construed; and that in fact many of the actions we perform – such as answering the telephone by saying 'Hello?', crossing at the green, or helping the needy can be better explained without them: In most such cases, we reflexively do what comes most naturally, and this, in turn, depends on our upbringing, habits, and social conditioning, not on whatever desires we may or may not suppose ourselves to have. Agents often act in accordance with norms or principles and without the intervention of desire, then, if those norms have been sufficiently internalized in the normal process of socialization so as to dispose them reflexively to such action when actualizing conditions obtain.

By a *norm*, I shall mean a recommendation, principle, rule, or law that prescribes behavior in the service of some favored goal; call such behavior *purposive.* The goal in question may be the achievement of some valued end-state, or it may be adherence to some valued standard of behavior. Conscious intentional behavior is *norm-governed* if it is caused by a disposition, normatively instilled in the process of socialization, to respond purposively to stimili under actualizing circumstances as the norm perscribes. By a *disposition* I mean a settled and regular tendency to behave in a certain way under certain recurrent kinds of circumstances (rather than an entity's structural propensity to react nomologically to certain kinds of causal-counterfactual conditions, even if those conditions should never obtain). A disposition is *normatively instilled* by such processes if there are social or physical factors in the environment that positively reinforce that response under its actualizing circumstances, and negatively reinforce its absence. Thus we can think of the social processes by which normative dispositions are instilled as not unlike the process Aristotle describes as *habituation*:[25] We learn to mimic repeatedly, under similar circumstances, the like behavior of elders or peers with whom we identify, or whose approval we seek; and the more frequently we rehearse the behavior under appropriate circumstances and are socially reinforced for doing so, the more natural and reflexive it becomes.

I shall refer to norms that govern the behavior of the Kantian self as *motivationally effective* norms. This does *not* mean that we must consciously strive to conform to these norms in order to be motivated by them. Nor need these norms be actually stated in a prescriptive form in order to be motivationally effective: We can easily imagine a community that adheres effortless-

ly and unselfconsciously to the norms that govern it, as Kant's fully rational beings do,[26] rather than agonizing over them as we often do. Rather, a motivationally effective norm is one that has been selectively instilled in the ways already suggested, such that we are ordinarily disposed to conform our behavior to it.

One mark that distinguishes us from other norm-governed sentient species, on this conception, is the centrality of shared, motivationally effective cognitive and linguistic norms that enable us to conceptualize all our behavior and experience to ourselves. Hence we are not merely norm-governed. We are governed by norms that enable us to know that we are. Thus the norms definitive of the Kantian self include, first and foremost, norms of *cognitive* behavior, i.e. prescriptive principles in accordance with which we are disposed to make sense of our experiences by generalizing over them and identifying particular experiences as instances of more general concepts: On this view, our thinking is ordinarily norm-governed. Secondly, these norms include norms of *linguistic* behavior, i.e. prescriptive principles in accordance with which we are disposed to apply to our particular experiences the general linguistic terms that symbolize the concepts we form. Thus our concepts and linguistic practices are similarly norm-governed according to this conception. Thirdly, the Kantian self is defined by norms of *emotional and gross physical* behavior, i.e. general prescriptive principles to which we are usually disposed to conform our emotions, actions, and habits as instances.

However, on this view, we no more have direct access to emotions, actions, or habits, unmediated by the norm-governed concepts in terms of which we make sense of them, than we do external events in the world at large.[27] Both internal and external phenomena are subject to interpretation by cognitive and linguistic norms. Thus, for example, if motivationally effective linguistic norms prescribe the vocabulary of desire to conceptualize motivation, a person will be disposed to use that vocabulary in interpreting their own behavior. On the other hand, if the vocabulary of desire is not prevalent, and altruistic behavior, say, is a motivationally effective social norm, as among the Zuñi of New Mexico, a person may be correctly described as moved by the perception of distress to render aid. By identifying the Kantian self primarily with its contextually determined cognitive and linguistic norms, rather than with the brute psychological phenomena interpreted as those norms prescribe, we leave open the question of what kinds of internal states should be invoked to explain the motivations of differently socialized selves, and how those states are to be conceptualized.[28]

In addition to the motivationally effective norms that govern the actual behavior of the Kantian self, there are also those norms with which the self actively identifies, and which constitute its *normative self-conception*. These may include, be identical with, or entirely disjoint from the motivationally effective ones. Thus a normative self-conception is related to the Kantian self as is a desired self-conception to the Humean self. Both evaluate our lower-order dispositions, beliefs, impulses, and goals as either conforming to or violating that self-conception, and both supply a *prima facie* motive for action.[29]

Now let us turn to the structure of the Kantian self. Like the Humean self, the Kantian self is also structured by rationality principles. However, each gives a different priority to the role of theoretical reason. On the Humean conception, reason and understanding have subordinate and instrumental roles. They enable us to organize and rank our desires, and to formulate maximally efficient plans for satisfying them. Hence they address only the strategic issues raised by what I shall describe as the *gross phenomena of action*: our consciously envisioned ends, our choices and plans, and the sequence of steps by which we carry them out. On the Kantian conception of the self, by contrast, the gross phenomena of action are only one kind of purposive behavior among many others, all of which are governed by motivationally effective norms, but *not* all of which are oriented towards the maximization of utility (witness the prohibition against eating one's peas with a knife). And so the principles of instrumental reason constitute only one kind of motivationally effective norm among many others.

On the Kantian conception, all such norms are themselves subordinate to those cognitive norms of generalization and concept-formation just mentioned. For on this view, the disposition to render all our experiences, including our experiences of our own conscious behavior, rationally intelligible is overriding. An experience is *rationally intelligible* if it can be identified by the agent as an instance of more general, motivationally effective norm-governed concepts. Consider the following example. I can make the most recent sound event I've experienced involving drums, bass, sax, and lead guitar rationally intelligible by identifying it as an instance of the norm-governed concept, 'Fusion', in part because this concept instantiates the more general norm-governed concept, 'Jazz', and 'Rhythm and Blues', which in turn instantiate the norm-governed concept, 'Music', all of which are motivationally effective for me. The concept, 'Fusion' is *norm-governed* in that its use is prescribed by

socially operative norms of language and musicology such as the following: 'Apply the concept, "Fusion" to sound events utilizing drums, bass, sax, lead guitar, blues, scales, polyrhythms, melodic improvisation, and a 4/4 meter'. The concept, 'Fusion' is *motivationally effective* for me if I am disposed to use that concept correctly, and respond to instances of it appropriately (i.e. with a range of positive or negative judgments and responses that recognize it as an instance of a certain kind of music, embedded in a certain cultural, political, and aesthetic context, and so on). If an initially unfamiliar event or experience could not be made rationally intelligible in this way, relative to our background assumptions, it is not clear how it could be integrated into the unified continuum of our experiences at all. Seemingly, it would stand apart as an unintelligible phenomenon to which the concepts we normally invoke to make sense of things bore no relation. But if it were in theory impossible for us to integrate it conceptually with our other experiences, it would seem equally impossible for it to constitute part of a unified self. I leave further elaboration of this point to Kant.[30] Similarly, if I behave in a way that cannot be identified as an instance of those more general norms of socially acceptable behavior which are motivationally effective for me, I may have trouble making sense of what I did, and why.

The structural relations among norms that are motivationally effective in the Kantian self are determined, then, by the disposition to organize experience in accordance with the norms of theoretical reason, i.e. by our disposition to individuate, compare, differentiate, and generalize consistently over different classes of experience to increasing degrees of inclusiveness. Hence, the Kantian self consists not of first- and second-order phenomena, but rather in a potentially infinite plethora of lower- and higher-order norms of increasing generality and comprehensiveness. The more comprehensive are our conceptualizations of our experience, the more internally integrated the self becomes. Thus preservation of the unity of the self and preservation of its rational intelligibility are equivalent.[31] I shall express this idea by ascribing to the Kantian self a *highest-order disposition to rationality*. For reason, on this view, supplies the primary and constitutive conditions of the internal coherence of the self, not just the instrumental conditional of its satisfaction.

Now in fact we experience the failure to achieve thoroughgoing rational intelligibility all the time. There are many events in the world of which we have trouble making sense, and often our own behavior is equally mysterious to us. So to ascribe to the Kantian conception of the self a highest-order

disposition to rationality is not to claim that we regularly succeed in rendering our experiences rationally intelligible. Further evidence for the view that we have such a disposition must be culled from indirect sources, and in the next section I shall try to provide some.

Earlier I claimed that the Humean conception of the self was future-oriented, heteronymous, and individualistic. By contrast, the Kantian self is present-oriented, autonomous, and social. It is *present-oriented* in that the self finds expression through actions that conform to the normative principles which presently govern it, not through realizing the envisioned state of affairs at which it aims. So, for example, if the norm 'Render aid to the needy' is motivationally effective for me, then I express myself by rendering aid to the needy, not by formulating and satisfying in the future my present desire to render aid to the needy. It has already been suggested that there are myriad actions we perform which are intentional, but actualize only present normative dispositions, not future objects of desire.

Second, the Kantian self is *autonomous* rather than heteronymous. For the conditions of its expression – i.e., the internalized social norms with which it is identified – are objects or states of affairs which are internal to the very constitution of the self. Actions determined by such normative dispositions express the self in virtue of their motivational source, not their actual or expected consequences. We do not normally await the outcome of our actions in order to decide whether we have successfully given vent to our impulses. Rather, we act (or, more often, react) in characteristic ways, determined by personality and circumstance, and hope for the best. The self is expressed in action, not in that for the sake of which it acts.

Finally, the Kantian conception of the self is *social* rather than individualistic. For if the self is to be identified with motivationally effective norms, then it is in fact defined by the particular social imperatives, recognized or unrecognized, to which it actively responds. It is not ultimately defined by a context-independent drive to achieve private satisfactions. This is merely the way *our* social norms make it look *to us*.

IV. RATIONALITY AND SELF-PRESERVATION

Next I should like to amplify and defend the claim that actual selves have a highest-order disposition to rationality, and so that the Kantian conception of the self is the correct one. I shall suggest one primary criterion of rational

intelligibility that *any* rational norm which is motivationally effective for us must satisfy, and three compensatory self-protective mechanisms we typically deploy – in vain – in order to preserve the rational coherence of the self against the threat of disunity.[32] There are other interesting constraints on rationality that can be derived from this primary one, but I shall not discuss them in this context.[33] Instead, I shall describe how our highest-order disposition to rationality enables us to solve the problem of moral paralysis in practice.

The primary requirement on rational norms is that they be *internally coherent*. This requires that the various components of our norm-governed experience be integrated and unified under the rubric of more general and comprehensive norms in the manner already described. This in turn requires not only that such norms satisfy the law of noncontradiction, i.e. that they be *consistent*. It also requires that they share features in common that allow us to apply to them more general norms which are motivationally effective for us, i.e. that we be able to *generalize meaningfully* over them.

For example, take the relatively general and motivationally effective cognitive norm that directs us to understand an external event in the world by seeking out its causal relations. Acting on this norm is logically consistent with that of trying to understand internal mental events, such as beliefs and feelings, by seeking out their causal origins in our upbringing, social environment, and previous experiences. But it is also similar in its reliance on causal explanation. The more general norm under which both are subsumed directs us to understand all the phenomena of experience by seeking out their causal connections.

However, we experience difficulty in applying this norm to all cases, and then we must resort to pseudorational mechanisms. For instance, the microphenomena studied by quantum physics seem peculiarly resistant to causal explanation, and our instinctive response to this fact is illuminating. We begin by *denying* the phenomenon, and cast about for flaws in the experimental design or apparatus to account for the apparent illusion. The intractability of the phenomenon to our attempts to wish it away are then met by a *rationalization*: We argue that there *must* be a causal explanation of this phenomenon, but that we are insufficiently equipped to discover its causes. When the evidence indicates the untenability of this position, we shrug our shoulders and proceed to *dissociate* the phenomena of quantum physics from the comprehensible world of causal relations we aspire to grasp. And we

thereby suffer the perplexity of trying, and failing, to see how the principles of quantum physics might be made to fit with everything else we think we know.[34]

Thus we defend the rational coherence of our experience by rationalizing, dissociating, or denying any phenomenon that threatens it. In *rationalization*, we apply a concept too broadly, ignoring or minimizing properties of that phenomenon that resist this generalization, and magnifying properties that support it. In *dissociation*, we resist or reject applicable generalization, instead relegating the phenomenon in question to the status of an alien and inscrutable enigma. In *denial*, we simply ignore or deny the existence of the phenomenon altogether, in order to maintain the appearance of conceptual coherence – or, as I shall say, the *pseudocoherence* – of our experiences. What makes these mechanisms *pseudorational* is that they each truncate or distort our experience in order to preserve its rational intelligibility. We can think of these three defense mechanisms, then, as ways in which our theoretical reason rallies, valiantly but ineffectively, to the challenge posed by conceptually unmanageble realities.

Consider next a comparable example of norm-governed emotional behavior. We are socially and biologically disposed to delight in the esteem or admiration of a person we love. We are similarly disposed to feel self-confidence and optimism upon receiving praise from some superior whose authority we respect. The more general, motivationally effective norm of which both of these are instances prescribes a positive joyful response to obtaining approval from someone whose regard is valuable to us.

However, we do not always respond emotionally in the way we recognize as appropriate. Suppose a highly valued personage in one's life – a respected colleague, say – shares too many extrinsic traits in common with other individuals one has valued highly in the past who have responded negatively to one's quest for approval. Suppose, for example, that she resembles one's mother, hated sibling, or former spouse. Then one may respond to her esteem or praise, sought-after and highly valued as it clearly is, not with delight or self-confidence, but instead with rage, resentment, or the suspicion of ridicule. One's awareness that such emotions are inappropriate may then lead one to *deny or suppress* the feelings in question, or to refuse to identify them for what they really are. Thus one may express one's resentment in the form of sarcasm or verbal abuse, and claim, upon being confronted, that one was only joking, meant no harm, that one's victim is oversensitive or insecure, and

so on. Alternately, one may *rationalize* one's anger by calling attention to the person's irritating imperfections, and claiming, for example, that anyone who speaks in a high whine, has dandruff, and wears galoshes all the time is bound to provoke blind fury, no matter what her virtues. Finally, one may simply *dissociate or disown* one's inappropriate emotional response by claiming that it overtook one as a blind, irresistable impulse, and was completely outside one's ability to control. People who take this last tack tend not to recognize the inconsistency involved in then promising that it will never happen again.

Similar considerations apply, finally, to the gross phenomena of action. Suppose, for example, that I conceive myself as a fair, tolerant, and sympathetic individual, and that most of my actions square with this normative self-conception: I am in fact loyal to my friends, actively concerned to promote others' wellbeing, and so on. However, I also spread unfounded and damaging gossip about individuals I dislike, thereby causing them severe personal and professional distress. This behavior would seem to be a patent instance of motivationally effective norms that are inconsistent with those governing the rest of my conduct, and so violate my normative self-conception. My disposition to preserve the internal coherence of my self-conception may then lead me to employ one of the defensive strategies just enumerated. First I may begin by *denying*, perhaps sincerely, that I behaved in this way at all; or recall the behavior but deny that it is an instance of spreading unfounded and damaging gossip. Rather, I may argue, it is merely an instance of indulging confidentially in harmless speculation. I thereby deny as well the very real damaging consequences of my behavior, and ultimately my own responsibility for bringing them about. Second, I may *rationalize* my conduct by arguing, say, that everyone gossips without thereby victimizing their subjects; and that after all, no one need worry who has nothing to hide. Thus the implicit thesis is that anyone who is damaged by unfounded gossip must have deserved it. Finally, I may *disown* or *dissociate* my behavior from that constellation of motivationally effective norms I identify as myself. By pleading that I am neurotic and easily threatened by others, and that mobilizing a network of social condemnation against them is a self-defensive reflex over which I have no control, I locate the cause of my behavior outside the scope of my voluntary agency.[35]

These self-defensive mechanisms for resolving internal incoherences are just as inadequate to integrate anomalies in our normative self-conception as they were to integrate anomalies in our normative emotional behaviour, and

in our normative conception of the physical world. They put a strain on the self that forces it to engage in yet more elaborate and irrational attempts to preserve its pseudocoherence, as, for example, when I conclude from the phenomenon of quantum mechanics that all events must be random and all regularities illusory; or when I attempt to cultivate an attitude of emotional indifference towards anyone whose approval I in fact value highly; or when I ascribe to the person I have maligned through gossip a malevolent power to make me feel guilty. These responses to the internal incoherence of the self are irrational because they themselves ramify that incoherence yet more widely throughout the structure of the self, and demand yet more elaborate attempts to ameliorate it; attempts which are similarly doomed to failure. The threat of ego disunity thus generates a stance of vigilant, self-protective defensiveness. For the more incoherent and irrational the behavior of the self, the more vulnerable to such threats it becomes.

For an imperfect but unimpeded Kantian self, acknowledging one's delinquent behavior as irrational is the best strategy for preserving the self against radical disunity, for this is to recognize that behavior as the painful threat it is to the rational coherence of the self. But since the Kantian self has, by hypothesis, a highest-order disposition to preserve the theoretically rational unity of its experience, the recognition that this unity is being destroyed by its own behavior disposes it, over the long term, to modify that behavior accordingly. In actual fact, it is questionable whether we ever truly succeed in reforming our conduct, without the prodding of these painful insights into our own irrationality.

However, even this is an option that not all selves are free to exploit. For though I have argued that all selves are in fact disposed to attain and preserve the internal coherence of their experience by the motivationally effective *cognitive* norm of rational intelligibility, it does not follow from this that all serves are governed by the *linguistic* norm prescribing correct use of the *concept* of rational intelligibility. Hence not all selves may be disposed to think of themselves as having a highest-order commitment to rational intelligibility *per se*, not to apply that concept correctly to their own behavior. And so the existence of demonstrably irrational behavior may not suffice to insure its rational modification. Perhaps one may believe, rather, that being a sensitive or virtuous individual, or being interesting, or politically committed, is more important than anything else. Then one will be impelled, under attack, to defend one's behavior at all costs in these terms, even in the face of

glaring inconsistencies, and regardless of the psychological discomfort it causes one to do so. Here one will be disposed to rationalize, dissociate, or deny any evidence that undetermines this defense. And of course this response itself will strongly indicate that those values did not, in fact, have primacy in one's hierarchy after all. For in this case, the defense of one's own behavior requires the suppression or distortion of one's values in the service of pseudo-rationality, and so *sacrifices* them for the appearence that one's behavior is rationally justified. And it is precisely the appearance of rationality that the self is, on this view, most centrally disposed to preserve. Any such values which are not finally consistent with the principles of theoretical rationality will be sacrificed similarly, in order to preserve the internal pseudocoherence of the self.

Thus do we resolve the problem of moral paralysis in practice. In fact, we are seldom torn by conflicting dispositions of the self, or inhibited from acting by uncertainty about our moral rectitude. More frequently, we simultaneously resolve the conflict and ensure our moral rectitude by appealing to some conceptualization of our actions that succeeds in preserving their coherence (or pseudocoherence) with the rest of our behavior, and thereby permits us to keep peace with our consciences. It is only to the extent that we fail recognizably to preserve coherence that we are led, by our instinct for self-preservation, to change our ways.

V. WHY I SHOULD NOT SPEND MY EVENINGS HOWLING AT THE MOON

The Kantian conception of the self outlined in this discussion treats the self as a natural phenomenon, comparable, in many respects, to other natural phenomena we encounter. Like the latter, it is causally determined and shaped by forces – biological, social, environmental – over which no one individual has any significant degree of control. As we do to other natural phenomena, we respond to the phenomenon of the self by trying to make it rationally intelligible to ourselves in socially conditioned, norm-governed terms. Like the failure of other natural phenomena, the failure of the self to conform to the norms by which we explain it provokes in us compensatory defense mechanisms, aimed at preserving the illusion of its rational intelligibility against the reality of its deviation. The failure of these mechanisms leads us to revise our thinking about the self, just as it does our thinking about the behavior of other natural phenomena, and to formulate alternative norm-

governed concepts to which the actual behavior of the self more closely corresponds.

But here the similarity with other natural phenomena ends. For unlike them, an essential feature – in my opinion, the *most* essential feature of the self is its very disposition to render its experience rationally intelligible. By contrast to our characterizations of the behavior of other phenomena that is conceptually anomalous, we are not let off the cognitive hook by dismissing our own behavior merely as, say, random rather than causal, or biologically deviant rather than stereotypical, or statistically improbable rather than likely. Instead, the failure of our defense mechanisms to sustain the appearance of rationality disposes us, in the case of the self, to recognize our behavior, specifically, as *irrational*, i.e., as incoherent and therefore a harbinger of ego-disintegration; and so to reform our behavior accordingly. Thus the self is unlike other natural phenomena in that its internal resources for altering its own behavioral patterns is identical to its tendency to understand them. And this tendency itself, which I have described as a disposition to rational coherence, in turn is identical to our disposition to literal self-preservation. This is what I mean by calling it a *highest-order* motivationally effective norm of human behavior.

Now this highest-order norm of theoretical rationality imposes an upper limit upon the proliferation of lower-order norms constitutive of the Kantian self, and so solves the problem of self-evaluation with which we began. For the ascent to $n+1$-order norms from which to evaluate the n-order dispositions and behavior of the self are finally subject to the requirement that all such $n+1$-order norms succeed in rendering those dispositions and behavior rationally intelligible in the sense explained. But to demonstrate their rational intelligibility is to provide an authoritative justification for maintaining them. For it answers the question of why we should behave in a certain way by demonstrating that it is in accord with the normative demands of theoretical rationality to do so. To then ask for reasons why we should do what it is demonstrably rational to do presupposes that in fact we should.

Thus there is in fact good reason why I should not spend my evenings howling at the moon, hence good reason why I should not desire intrinsically, at the highest order, to do so. This is that I have a certain norm-governed, coherent self-conception that includes a concept of what it means to be and to behave like a human being, with which howling at the moon is inconsistent. This concept is motivationally effective for me in that it disposes me to pick

out, correctly identify, and evaluate instances of characteristically human behavior as such, to form justified expectations about my own and other people's behavior in light of it, and unreflectively to conform my own behavior to it.[36] Of course, like most human beings, I have the *capacity* to violate this concept in my own behavior; but if I am sufficiently well socialized, I lack the *disposition* to do so. To then spend my evenings howling at the moon despite this would be to violate my own rationally intelligible self-conception, i.e. my conception of the kind of creature I am. It would force me to deny, rationalize or dissociate myself from my own behavior, in order to preserve my sense of self as a human being.

But these self-defensive strategies would probably fail. I could not for long *deny* or ignore the fact that I regularly spent my evenings howling at the moon, without provoking all the attendant difficulites that anmesia tends to bring. And to what *rationalization* could I appeal to restore intelligibility to my conception of what I was doing: that everyone has their little idiosyncracies, perhaps? This appeal would certainly fail, since as a matter of empirical fact, the range of behavior we are willing to recognize under the aegis of 'human idiosyncrasy' simply does not extend this far. Of course our conception of human nature responds flexibly to the variety of circumstances and ways in which human nature develops. Nevertheless, it is sufficiently circumscribed so that we are disposed to recognize a genuine anomaly when we encounter it. That is, we are disposed to differentiate such behavior from our norm-governed concept of how human beings are characteristically disposed to behave. And so I, as the anomalous agent, would be self-defensively moved to *dissociate* my own identity as a human being from the actual behavior I performed. And then I could retain my sense of humanity only by disavowing my own agency; or retain my agency by disaffiliating my connection with humanity. That I would in either case effect such a radical incoherence within the self is why it would be irrational for me to spend my evenings howling at the moon.

However, the perspective of rational intelligibility from which we are disposed to survey, evaluate, and organize the lower-order normative components of the self may not be the perspective of our explicit normative self-conception. For if we are without illusions about the degree of rationality we are in fact able to attain, we may disavow any conscious commitment to rationality whatsoever. This may lead us, in turn, to reject the rational perspective as impersonal, and detached from everything that gives our lives

meaning. But I am inclined to dismiss this stance, too, as a bit of self-deception that is ultimately incoherent. For without a commitment to rationality, however involuntary it may be, our lives could literally have no meaning in any sense of the word; and in practice we are forced to recognize this. A failure of rational intelligibility is a failure of comprehension; a lacuna in our accounts of ourselves, other people, and the world at large. A failure of comprehension in turn signals our irradicable alienation from the object under scrutiny, i.e., the admission of the opaque, the incoherent, the inexplicable, into our conception of reality; and this conflicts with our most basic instinct of self-preservation. For typically constituted human beings, the disintegration of the self is psychologically equivalent to the death of the self, and this is a state we are disposed to avoid at all costs. The Kantian conception of the self acknowledges this important fact about us. On this conception, then, a rational self is a fully unified and integrated self; a self to which, I have tried to show, human beings are characteristically disposed to aspire.

ACKNOWLEDGEMENT

The writing of this paper was supported by an Andrew Mellon Postdoctoral Fellowship, Stanford University. Earlier versions of it were presented to the Philosophy and Anthropology Group and the Department of Philosophy, both of the University of Michigan; and the Departments of Philosophy at Stanford, U. C. Berkeley, the University of Minnesota, and the University of Pennsylvania. I have benefitted enormously from comments received on those occasions. I should also like to thank Michael Bratman, Jeffrey Evans and Allan Gibbard for their criticisms of earlier drafts of this paper, and Michael Markert for his contributions to the dissertation chapter version of the same title (Chapter II of 'A new model of rationality', Ph.D. diss., Harvard University, 1981).

NOTES

[1] This conception is probably not embraced in its entirety by any one of its adherents. Rather, different facets of it are called into service to do different philosophical jobs: to explain behavior, for example; or to analyze moral motivation, or freedom of the will. Thus the picture I shall sketch is a composite one, drawn from many different sources.

[2] Richard Brandt and Jaegwon Kim, 'Wants as explanations of action', in N. S. Care and C. Landesman (eds.), *Readings in the Theory of Action* (Indiana University Press, Bloomington, IN, 1969), pp. 199–213.

Brandt and Kim explicitly mean to construe wants (or desires) as theoretical constructs, with no experiential analogues (pp. 200–202 and fn. 2). This interpretation allows them to apply the concept of a want or desire to the explanation of a broader range of behavior than would be suggested by the ordinary sense. However, five of their six proposed criteria for the correct usage of 'x wants p' make explicit references to x's

experience of such feelings as joy or disappointment in the attainment or nonattainment of p, pleasure in entertaining the thought of p or in the occurrence of p, and an impulse to do the act that x believes will eventuate in p. To analyze the concept of a want or desire for p in terms of joy or pleasure at the satisfaction of that want and a felt impulse to achieve that satisfaction seems inconsistent with denying that 'want' denotes an experience. If it denotes a constellation of experiences then presumably it denotes each conjointly in that constellation. My quarrel here is not with Brandt's and Kim's analysis, but rather with their attempt to divest the concept of a want or a desire of the particular experience (or conjunction of experiences) that individuate it from other motivational states.

[3] Bernard Williams, 'Persons, character and morality', in A. O. Rorty, Ed., *The Identities of Persons* (Berkeley, University of California Press, 1976).

[4] This distinction is first made explicitly in Harry Frankfurt's seminal article 'Freedom of the will and the concept of a person', *Journal of Philosophy LXVIII*, No. 1 (January 1971), pp. 5–20. Frankfurt's main thesis is comparable to Wright Neely's apparently independent treatment in 'Freedom and desire', *Philosophical Review LXXXIII*, No. 1 (January 1974), pp. 32–54.

Although Neely emphasizes the contrast between the ordinary sense of 'desire' as one motive to action among many and the extended philosophical sense that includes all such motives to action, he makes it equally clear that the advantage of the philosophical sense is that it implies means for analyzing all the multifarious motives for the action in terms of 'desire' in something like the ordinary sense. Thus he seems to mean 'desire' in the technical sense to *cover or substitute for* duties, purposes, intentions, and volitions – as would similarly technical terms like 'conation' and 'appetition', each of which could be interpreted or analyzed in terms of 'desire' in a more ordinary sense, as Neely's examples of duty and wellbeing illustrate. If this interpretation can be carried through, such that each motive to action can be claimed to include a desire in the ordinary sense, then using 'desire' in the technical sense to denote all such motives has obvious advantages over terms like 'conation' and 'pro-attitude'.

[5] This conception of the Humean self as structured by the principles of instrumental rationality is explicated in greatest detail in Chapter VII, 'Goodness as rationality', of John Rawls' *A Theory of Justice* (Cambridge, MA, Harvard University Press, 1973). See especially Sections 63–64 and the bibliography cited there.

[6] Hume, *A Treatise of Human Nature*, ed. L. A. Selby-Bigge, (London, Oxford University Press, 1968), Book II, p. 415.

[7] The classic statements of the belief-desire model of action are to be found in Richard Brandt and Jaegwon Kim, 'Wants as explanations of action', (*op. cit.*) and Donald Davidson, 'Actions, reasons and causes', in Care and Landesman, pp. 179–198, from which Neely's use of 'pro-attitude' stems.

For applications in moral philosophy, see, for example, Phillippa Foot, 'Reasons for action and desire', *Aristotelian Society*, Supplementary Volume XLV (1972), pp. 180–210 and 'Morality as a system of hypothetical imperatives', *Philosophical Review* 81 (1972), pp. 305–316.

[8] See Note 3.

[9] An example of this strategy is to be found in Richard Brandt's 'Traits of character: A conceptual analysis', *American Philosophical Quarterly 7*, No. 1 (January 1970), especially 27–30. This analysis builds on the earlier paper by Brandt and Kim (see Note 2).

[10] Immanuel Kant, *Groundwork of the Metaphysic of Morals*, trans. H. J. Paton (Harper Torchbooks, New York, 1964), Ac. 441–5.

[11] "This relation, whether based on inclination or on rational ideas, can give rise only to hypothetical imperatives: 'I ought to do something *because I will something else*.'" *Ibid.*, Ac. 441; italics in original.

[12] Bernard Williams argues this point in 'Egoism and altruism', in *Problems of the Self* (Cambridge University Press, New York, 1975).
[13] *Op. cit.*, Note 4. Also see Gary Watson, 'Free agency', *Journal of Philosophy LXXII*, No. 8 (April 1975), pp. 205–220.
[14] Frankfurt, p. 16, (*op. cit.*, Note 4).
[15] In 'Responsibility for self', in Rorty.
[16] *Op. cit.*, Note 13.
[17] *Ibid.*, p. 215.
[18] *Ibid.*, p. 213.
[19] Wright Neely makes this point in anticipation of Watson's analysis (*op. cit.* Note 4), p. 42. Watson does not use the term "standpoint" as I do the term "self-conception". He means "the point of view from which one judges the world". (p. 216) He doubts the validity of the picture of the Humean self as scrutinizing and evaluating the worth of its first-order desires. Rather, he believes that

> [agents need not usually] ask themselves which of their desires they want to be effective in action; they ask themselves which course of action is most worth pursuing. The initial practical question is about courses of action and not about themselves.

But this seems part of a general plan to throw out the baby with the bathwater. For in denying that we evaluate our first-order desires from the perspective of second-order ones, he seems to want to deny that we act self-consciously at all. But surely one consideration that favors any action we deem worth performing is that it is consistent with actions performed by the kind of person we aspire to be. The 'point of view from which one judges the world' is the point of view of a certain kind of self whose capacities for critical scrutiny are exercised as often on itself as on other objects.
[20] Bernard Williams, 'A critique of utilitarianism', in J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against* (New York, Cambridge University Press, 1973). Also see Note 3. W. D. Falk makes a similar point in 'Morality, self, and others', in Judith J. Thompson and Gerald Dworkin, eds., *Ethics* (New York, Harper and Row, 1958).
[21] This point about the character of intentional action without agency is found in Frankfurt (*op. cit.* Note 4), p. 16.
[22] See particularly A67=B92 – A136=B175 in the Analytic, and A299=B355 – A338=B396 in the Dialectic of Immanuel Kant, *The Critique of Pure Reason*, trans. Norman Kemp-Smith (New York, St. Martin's Press, 1970). My indebtedness is predicated on a reading of Kant with which some may disagree, i.e., that the alleged differences between the forms of intuition, the understanding, and reason are in fact (despite Kant's frequent caveats to the contrary) not differences in kind or 'faculty', but rather differences in degree of generalizing capacity, of which the categorical imperative is an instance applied to the special case of action. Those who do not find this reading prima facie suggestive or plausible should simply ignore my frequent appeals to Kant's authority in this discussion.
[23] See my 'Kant's First- and Third-Person Criteria of Humanity' (unpublished paper, 1981).
[24] 'The rationality of military service', in *Conscripts and Volunteers: Military Requirements, Social Justice, and the All-Volunteer Force* (Maryland Studies in Public Philosophy; Totowa, NJ, Rowman and Allenheld, 1983), Chapter III, 8.
[25] Aristotle, *Nicomachean Ethics*, trans. W. D. Ross, in Richard McKeon, (ed.), *Introduction to Aristotle* (New York, The Modern Library, 1947), Book II.
[26] See Note 10, *op. cit.* Chapter II.
[27] "Even as regards himself – so far as man is acquainted with himself by inner sensation – he cannot claim to know what he is in himself. For since he does not, so to say, make himself, and since he acquires his concept of self not a priori but empirically, it is

natural that even about himself he should get information through sense – that is, through inner sense – and consequently only through the mere appearance of his own nature and through the way in which his consciousness is affected." – *Groundwork*, Ac. 451–2, *op. cit.* Note 10. Also see *Critique of Pure Reason, op. cit.* Note 22, B152–159 and A 551, n.

[28] I have been greatly encouraged by the responses of some anthropologists to this idea. Thanks are particularly due to Sherry Ortner, Kit Roberts, and Michael Taussig for their comments, insights, and wealth of ethnographic confirmation.

[29] Although not necessarily a prima facie *purpose of* action (see my 'Narcissism and moral alienation', unpublished paper, 1984).

[30] Kant, of course, would claim that these requirements must be satisfied in order for us to have conscious experience at all. See *The Critique of Pure Reason, op. cit.* Note 22, e.g., A94, 99–103, 108, 111, 116, 121–2, B132, 141–3, 164. *passim*. In the most extreme case, a complete failure of integration of experience would result in what Kant would describe as an "unsynthesized manifold", i.e., a succession of perceptions disunified and discontinuous from one moment to the next, hence giving rise to no psychological sense of self at all.

[31] "All our knowledge starts with the senses, proceeds from thence to understanding, and ends with reason, beyond which there is no higher faculty to be found in us for *elaborating the matter of intuition and bringing it under the highest unity* of thought." (A298 = B355; my emphasis) ... "From this we see that in inference reason endeavors to reduce the varied and manifold knowledge obtained through the understanding to the smallest number of principles (universal conditions) and thereby to achieve in it the highest possible unity". (A305) *Critique of Pure Reason, op. cit* Note 22. Also compare A300 = B359.

[32] The self-defensive mechanisms I discuss in the following pages seem to be as universal and innate to human mental functioning as is the highest-order disposition to rationality. Robin Horton's discussion of the "'closed' and 'open' predicaments" in his article, 'African traditional thought and Western science' (in Bryan R. Wilson, ed., *Rationality*, New York, Harper and Row, 1970) sheds considerable illumination on this phenomenon from a cross-cultural perspective.

[33] See my *Rationality and the Structure of the Self*, Chapter VI (unpublished manuscript, 1984).

[34] Compare Thomas Kuhn's discussion (in *The Structure of Scientific Revolutions* (Chicago, The University of Chicago Press, 1970)) of scientists' responses to anomaly in scientific theories, especially pp. 62–6, 78.

[35] This overly abbreviated discussion of dissociation has profitted from Harry Frankfurt's 'Identification and externality', in Rorty, and from John Wilson's 'Freedom and compulsion', *Mind 67* (1958), pp. 60–29, although I am not in final agreement with much of what they have to say.

[36] Cf. Kant: "... no single creature in the conditions of its individual existence coincides with the idea of what is most perfect in its kind – just as little as does any human being with the idea of humanity, which he yet carries in his soul as the archetype of his actions ..." *The Critique of Pure Reason, op. cit.* Note 22, A318.

Department of Philosophy,
University of Michigan,
Ann Arbor, MI 48109,
U.S.A.