What Do Expressions of Preference Express?

I—Actions, Reasons, and Preferences

Some, but not all, of our behavior deserves to be called 'action'. We distinguish among *our doings* in a broad sense, a special class of performances that are both *doings* and (therefore) *ours* in a richer and more demanding sense. Action is behavior that is rational, in the sense that the question of what *reasons* can be given for actions is always at least in principle in order. Actions are performances that are caught up in our practices of giving and asking for reasons as moves for which reasons can be proffered and sought. Although there may be much more to the concept of action than is captured in this characterization, the connection between action and reasons is sufficiently tight that one could not count as understanding the concept of *action* (as even minimally mastering the use of that and cognate words) unless one also counted as in the same sense understanding the concept of *reasons* (for action).

One specifies a potential reason for an action by associating with the performance a goal or an end: a kind of state of affairs at which one understands it as *aiming*, in the sense that its *success* or *failure* is to be assessed accordingly as it does or does not bring about a state of affairs of that kind. Because this is the form of reasons for action, actions, as essentially performances for which reasons can be offered or demanded, are also essentially performances whose success or failure can be assessed.[1] Talk of there being *better reasons* for one action than another among some set of alternatives is convertible into talk of relative *values* associated with the ends, goals, or aims of those actions.

Enlightenment philosophers faced a problem integrating this constellation of concepts of agency with the concepts of the new physics, inaugurated by Galileo, pursued by Descartes, and in many ways perfected by Newton. Described in the language of the physics they developed, the world does not come with *values*—or, equivalently, with *reasons for action*—in it. In this specific sense, the mathematized language of the new physics *practically disenchants* the world it describes.[2]

The characteristic response of Enlightenment philosophers to this challenge is to seek to understand values and reasons for action as themselves products of human activity—as introduced into the physical world by our practices and attitudes. Talk of values is talk of

[1] Compare (and contrast) the relations between reasons for *judgments* and the liability of judgments to assessment as *true* or *false*.

[2] Of course a corresponding problem arises concerning the placement of reasons for *judgment*—and hence the pursuit of the new science itself—in the mathematized world described by physics. In a *tour de force* that brilliantly epitomized some of the deepest impulses of the Enlightenment, Spinoza in his *Ethics* sought to reduce the problem of practical reason to that of theoretical reason. What we have reason to *do*,

reasons for action; talk of reasons for action comes into play only in the context of the behavior of rational creatures. We institute values and reasons by doing what we do, including reasoning about what we do. Our rationality consists, in the end, in properly taking account of the values and reasons for action we ourselves have introduced into the world.

This master-thought of Enlightenment theories of practical rationality can be developed in different ways. My concerns in this essay are oriented by one great divide among them: the distinction between theories that take it that our reason-and-value-instituting activities can adequately be specified as such in a resolutely *nonnormative* vocabulary, by focusing on what agents do in fact choose to do, and those that insist on the contrary that only a normatively rich vocabulary—one making irreducible appeals to what agents *ought* to choose—will do to ground talk of reasons for action.

The first approach is animated by the thought that the only theoretical grip we have on what is valuable is the activity of valuing or treating as valuable. Agents sort possible ends or goals into more and less valued ones (take them to be more and less valuable) by pursuing some at the expense of others. We discover their reasons for action by discovering what they actually take as their ends. One cardinal culmination of the development of this tradition is the contemporary economic theory of rational choice. Pursuing one end at the expense of another is making a *choice*. Where the choices an agent is disposed to make hang together in the right way, they can be understood as

---

in the end, is to give the reasons and make the judgments that improve science, to perfect our knowledge

revealing *preferences* for some kinds of ends (features of outcomes) over others. For choices to 'hang together in the right way' to be understood as revealing preferences is for them to admit a measure, *utility*, such that in any particular situation one is disposed to choose whatever end *maximizes* that measure of preference. To be practically rational—not merely to *make* choices, but to have *reasons* for them—can then be understood as having dispositions to choose that admit of such a maximizing interpretation. Where such an interpretation—the conditions for which can be made quite precise— is available value can be identified with utility, the measure of preference that is maximized in choice.[3]

The second approach takes as its starting point the idea that there is an important distinction between what *causes* actions and choices, and what *justifies* or provides *reasons* for them. Neither mere inclinations or dispositions to choose, nor the preferences that under favorable circumstances can be seen to be revealed by them, by themselves provide reasons for those choices—though of course they can cause such choices. (As Anscombe reminds us, to explain one's action by saying "I just wanted to, that's all," is not to offer a reason for it.) Reasons for actions should be understood in terms of values that are in principle intelligible apart from consideration of what the agent in fact would choose to do. The wellsprings of rational action are found not in raw behaviorally revealed preferences, but in a richer notion of an agent's *endorsement* of an end. For an

---

of the physical world, and thereby to become the mind of God.

[3] In fact, what an agent chooses depends not only on what she wants (prefers, values), but also on what she *believes*—most importantly about the outcomes consequent upon various candidate choices. Where the idealization that these can be held fixed is removed, one must introduce into the interpretation a further parameter: the agent's subjective (conditional) *probabilities*. Then we will say that utility is what the agent is *attempting* to maximize, and that rational agents maximize *expected* utility. Relaxing the idealization in this direction does not yet amount to making the transition from thinking about the sorts of

agent to endorse an end (and so indirectly certain actions or choices) is to take it to be one she *ought* to pursue, to treat it as one she is *committed* or *obliged* to pursue, one that is *worthy* of being pursued. To be practically rational is to act according to the values one recognizes, the commitments, obligations, or duties one acknowledges as binding. This is the point Kant is making in defining a rational will as the capacity to derive acts from conceptions of *laws*.

The second sort of approach may be called 'minimal kantianism' about reasons for action. Kantians in this sense may accept the Enlightenment insight by insisting, as Kant himself did, that no norms (values, obligations, commitments, duties…) bind agents apart from the endorsement or acknowledgment of those norms by the agents themselves—that the normative statuses in question are not intelligible apart from reference to the normative attitudes of those who recognize them.[4] But they are committed to a construal of such endorsement or acknowledgment of the bindingness of norms that goes beyond mere dispositions to act in a certain way, as described in a vocabulary that eschews reference to values, obligations, commitments, or duties. The minimal kantian understands having a reason as acknowledging something that is there independently of the agent's acknowledgment of it, and accordingly understands choices as answering to norms that are not instituted by an agent's dispositions to choose. Being rational is being sensitive, in

---

parametric choice situations addressed by decision theory to thinking about the sorts of strategic choice situations addressed by game theory.

[4] It is 'minimal' kantianism because nothing in what has been so far ascribed to the approach requires its completion by some analogue of a categorical imperative. Kant may be thought to have moved too quickly from consideration of the way in which endorsement can swing free of inclination to the conclusion that inclination is simply irrelevant to what we have reason to do. This move can usefully be compared to the classical rationalists' inference from the claim that awareness consists in the application

one's practical reasoning, to the normative statuses (commitments, obligations) which, when acknowledged, provide the only real reasons for action.

The differences between the two sorts of approach can be subtle, but they are real. The main divide is over whether reasons for action are grounded justificatorily—not causally—ultimately in normative statuses (for instance, duties), so that only such statuses can serve as the source of reasons. On the minimal kantian view, relative value, codifying reasons for action, provides not only a *measure* of preference, but also a *standard* of preference.[5] That is, choices can legitimately be assessed as better or worse, depending on whether they express the acknowledgment of commitments, obligations, and so on that actually bind the individual. This is to accept a view of practical reasoning as essentially involving concern with what ends of action *ought* to be recognized.

In drawing this distinction I want to put to one side the issue of what reasons we could have to commit ourselves or acknowledge one obligation rather than another—something Kant himself is much concerned with. One might be inclined to insist that only commitments that were themselves rationally undertaken could serve as reasons for action. But it is instructive to consider things at the level of abstraction at which the minimal kantian *need* no more insist on the antecedent rationality of the individual reason-grounding normative statuses taken as the ultimate source of reasons for action than the orthodox economic rational choice theorist *need* insist on the antecedent rationality of the

---

of concepts, to the claim that concepts must owe nothing to conscious experience save the occasions of their application.

[5] *MBA* 25.

individual preferences or dispositions to choose taken as the ultimate source of reasons for action. I am suggesting that we think about that economic theory in relation to a view (picking up, to be sure, only one element of Kant's) according to which the not-necessarily-themselves-rational grounds of reasons for action are *commitments*, *endorsements*, or *obligations*. Like preferences construed as exhaustively manifested by dispositions to choice behavior, these can be treated as rationally (=inferentially) *articulated*, without being treated as, as a group, *grounded*. In each case, some of them may be something with which agents just find themselves.

II—Preference: Behavioral vs. Attitudinal, and Raw vs. Considered

In *Morals By Agreement*, David Gauthier develops the most powerful and significant variant of the classical economic account of rationality, within the broad outlines of the rational choice approach. His account agrees with that approach in identifying rationality with the maximization of a measure of preference, and thereby, he says, disclaiming all concern with the ends of action.[6] He accordingly disavows the minimal kantian approach to reasons for action distinguished above. But he insists as well that however useful it may be in economics, the classical theory—according to which preferences are simply read off of choices—is not adequate as an understanding of practical rationality in general.[7]

Gauthier begins his reconstructive enterprise at the very bottom: "To move from the economist's account to a view of rational choice adequate to understand rational behavior, we must begin by reconsidering the conception of preference."[8] The notion of preference he introduces in order to create the space for his sophisticated variant of the master idea of rationality as maximizing (acknowledged) value is broader than the one that figures in classical theories in two ways. First, he thinks of preferences as not only *revealed* in *choice behavior*, but as also *expressed verbally*. Second, he distinguishes the *raw* preferences that are manifested in these two ways from the *considered* preferences, stable under experience and reflection, to which they can give rise. Only the latter are understood as providing genuine reasons for action. These two moves are introduced in a few pages at the very beginning of the book. My main concern in this essay is to look more closely at the understanding of preference that is implicit in them. My overall thesis will be that there is a significant unacknowledged tension between this reconstrual of preferences and Gauthier's rejection of minimal kantianism in favor of approaches to practical reasoning that restrict it to a merely instrumental role. Since I think the reasons he advances for his rethinking of the concept of preference are compelling, this tension may require us to reassess Gauthier's own classification of the view he builds on that basis.

By distinguishing (merely) *behavioral* manifestations of preferences in dispositions to choice behavior from *attitudinal* manifestations of preference, in (merely) verbal

---

6 *MBA* 26.
7 *MBA* 27.
8 *MBA* 27.

behavior—expressing what one wants by *saying* what one prefers—Gauthier drives a wedge between preference and choice, both conceptually and operationally.[9] Doing that is the first step in distinguishing *value* from *utility*, which is what makes possible his eventual identification of practical rationality with the maximization of value (the measure of considered preference) rather than utility (the measure of raw, behavioral preference). These distinctions are needed to make room for Gauthier's notion of practical rationality as *constrained* maximization. For if value is definitionally identified with utility, utility is definitionally identified as the measure that is maximized in choice behavior, and preference as what utility is the measure of—as classical economic rational choice theories do—then Gauthier's approach is ruled out conceptually: it cannot then even be coherently formulated.

The intratheoretic need for taking seriously manifestations of preference not just in (dispositions to) choice behavior, but also in (dispositions to) verbal behavior is significant as well in the context of Gauthier's second move. For the process of consideration, by which one's raw preferences are transformed into final preferences, stable under experience and reflection, requires preferences to be expressible in a form in which they can serve as premises in reasoning about hypothetical situations. Only so can their interactions be gauged and calibrated, some rejected or modified in the light of others. An example would be reasoning of the form: "If I prefer outcome A to outcome B, and prefer outcome B to outcome C, then I ought not to prefer outcome C to outcome A." Only what is propositionally contentful, in the sense of verbally expressible in declarative

[9] *MBA* 28.

sentences, can figure as antecedents of conditionals (codifying premises of inferences) such as these. So the second move requires at least that preferences be verbally expressible.

The basic idea is to "accept the general explanatory schema: choice maximizes preference fulfillment given belief," while rejecting "the trivialization of this schema that results from denying independent evidential access to each of its terms—to choice, preference, and belief."[10] And this idea has good and sufficient pretheoretic motivation. It is not simply an *ad hoc* device to clear the way for the theory of rational interaction to follow. For one thing, the roles played by preferences and the role played by beliefs in practical reasoning preferences are in many ways symmetrical. But beliefs are manifested both in what we *do* and in what we *say*. Though these two sources of evidence concerning belief often are consilient, they can come apart (and when they do, they may lead to action that is irrational, and not merely hypocritical). Construing preference on the model of belief would dictate admitting two sources of evidence concerning the former, as well as the latter.[11]

Another presystematic reason to move in the direction Gauthier urges is this. If we think it is irrational in a basic sense to have preferences that do not permit the definition of a utility measure—for instance, because of dispositions to cyclical, intransitive choices of the sort ruled out by the conditional forwarded above—but want to treat individual

[10] *MBA* 30.

preferences themselves as data that are neither rational nor irrational, we need to take account of a process of assessing a whole *set* of preferences, each element in the light of its fellows, and all in the light of collateral beliefs about the relations among the outcomes over which those preferences are defined. We want to be able to criticize those disposed to choose A over B, B over C, and C over A as *irrational*. Yet if we define preference as that the measure of which (utility) is maximized in choice behavior, we are obliged to deny that there can be cyclic preferences. The agent in question must be characterized as simply not having any preferences in the vicinity of these choices; and it is not irrational not to have preferences. We are thus debarred on the classical economic theory from satisfying what would seem to be a cardinal criterion of explanatory adequacy for any theory of rationality.

Another such criterion of adequacy on theories of practical rationality, besides that of addressing what appear to be formally defective structures of preference, is dealing with cases of *akrasia*, or weakness of the will. Sometimes, it seems right to describe agents as acting against their own preferences: choosing to smoke, while insisting that they prefer health to illness, and believe that smoking leads to illness, and so on balance prefer not smoking to smoking. Definitionally tying preference to choice behavior obliges the theorist to explain these appearances away: weakness of the will cannot coherently be described as a form of irrationality, and so cannot pose problems for a theory of practical rationality. Acknowledging the category of verbally expressed preferences that may not

[11] Davidsonian interpretation insists on doing this: neither preference nor belief can be read off of nonverbal behavior, given their joint role in rationalizing that behavior. Not only must each be attributed as part of a story that includes an account of the other, but verbal behavior is crucial in attributing both.

be manifested also in dispositions to choice behavior is surely better than this Procrustean response. Then the akratic can coherently be described as having genuinely conflicting preferences, and is potentially assessable as acting irrationally. These are the sorts of considerations Gauthier has in mind when he says that however adequate the classical account of preference exclusively in terms of choice may be for the purposes of economic theory, it will not do as an idiom in which to formulate a more general theory of practical rationality.

III—What Do Expressions of Preference Express?

For the acknowledgment of attitudinal preferences, manifested in dispositions to verbal behavior, to do this sort of work in a general theory of practical reasoning, the notion of preference cannot be limited to what is *also* revealed in dispositions to choice behavior. That is, the move must not be limited simply to adding one more way preferences in that sense can be manifested. Both Gauthier's systematic aspirations, and each of the three presystematic reasons for admitting expressions of preference in verbal, as well as choice behavior canvassed above, requires that in at least some cases we can identify genuine preferences, even in the absence of dispositions to choice behavior, on the basis *just* of dispositions to verbal expressions of them. This is what is meant by saying that verbal expressions of preference can give us evidence about preferences that is in principle independent of (even dispositions regarding) choice behavior. Nor will it do to make a

purely formal assimilation of the latter to the former, on the basis that verbal behavior *is* choice behavior, namely choice of words to utter. For what matters is the *content* of the words—that they be expressions of preferences relating, in some cases (think of the akratic) to the same outcomes to which (merely) behaviorally revealed preferences are addressed. As Gauthier says, in some contexts "the two dimensions of preference are distinct, and may be in conflict."[12]

At this point, however, a question arises: What *are* verbal expressions of preference? If we had an independent grip on the notion of preference, we could leave this question to the semantic theorist—the one who tells us what verbal expressions of *anything* are. Thus if we start with dispositions to choice behavior, and define preference exclusively by appeal to such behavior (taking preference to be what is measured by utility, which is taken to be what is maximized in choice) we can then go on to specify the truth conditions of (third person) attributions and (first person) expressions or avowals of preferences in those terms. Similarly, if we thought, as William James sometimes seemed to have, that we could define belief behaviorally in terms of what agents are disposed to do, independently of what they are disposed to say, then we could go on to give truth conditions for verbal expressions of belief—claims—in those terms. Classical rational choice theory is the conative analog of this sort of cognitive pragmatism. As classical pragmatism had to assume that the preferences on which one acted could be held fixed, or at least independently determined, so classical rational choice theory has to assume that the beliefs on which one acts can be held fixed, or at least independently determined—

[12] *MBA* 28.

presumably in each case, verbally.[13] Once one relinquishes this division of labor, as Gauthier does on the conative side, the explanatory strategy that depends on it must be foregone as well. Preferences must now be understood as what can be expressed either in choice behavior or in verbal expressions of preference. Choice behavior, we may assume, is no worse off from an explanatory point of view than it was for the classical economic rational choice theorists. But now we need an independent grip on the notion of expressions of preference, in order to understand what preferences themselves are.

We cannot just say that expressions of preference are first person avowals whose main verb is a term like 'prefer', 'want', 'desire', or 'value'. For the hard question is what these terms *mean*, and (so) what must be true of a verb for it to be *like* them in the relevant sense. The main clue that we have is that the locutions we are after, while they sometimes express preferences that are not also manifested in choice behavior, in the central and predominant cases express preferences that are *also* manifested in choice behavior. As Gauthier says: "We assume the coincidence of revealed and expressed preference unless we have clear evidence of divergence."[14] Thus in common, central, and favored cases, expressions of preference will be true, if true, because of their relation to preferences intelligible also in terms of dispositions to choice behavior. But what we want to understand is Gauthier's innovation: the purely attitudinal preferences that are *only* manifested in dispositions to verbal expression, and *not* also in dispositions to choice

[13] Of course one can interpret agents by simultaneously assigning subjective probability functions and subjective preference functions, as David Lewis does. The present point concerns not the relation between attributions of preference and attributions of belief, but the division of labor between behavioral evidence and verbal evidence. The claim is that classical pragmatism about belief and classical rational choice theory about preference line these distinctions up.

behavior. To understand the surplus of expressed over revealed preferences, we need to know what can make it correct to utter an expression of preference that does *not* correspond to a preference that is also behaviorally revealed.

To focus the question, suppose that in some language (perhaps our own), we find two locutions, each of which can correctly be used to report the presence in oneself of preferences corresponding to dispositions to choose, but each of which can also be used properly under other circumstances. Suppose further that they differ in what those other circumstances are. What then would make (at most) one of them a genuine preference-expressing locution, and (at least) the other a merely disjunctive locution, expressing preferring-or-Φing A to B (where Φing might be, say, admiring more, or finding more dramatic)? If an alien language had only one comparative verb properly used to report dispositions to choice behavior, but that was also sometimes properly applied in their absence, what would settle whether (not epistemically, but in the sense of "make it the case that") *all* its uses expressed genuine preferences, some behavioral and some merely attitudinal?

I don't think Gauthier answers this question. But if he does not—if he does not tell us what verbal expressions of preference are—then he does not tell us what preferences are either. For his abandonment of the classical definition of preference in terms of choice behavior means that for him, in the crucial surplus cases where attitudinal preferences outrun behavioral ones, preferences just are whatever verbal expressions of preference are

[14] *MBA* 28.

expressions of. Gauthier should be thought of as introducing a new primitive: verbal-expressions-of-preference, and extending the classical choice-behavioral concept of preference to include states picked out by their relation to this sort of locution. The challenge I am highlighting is to say more about the speech acts or linguistic expressions invoked by the new primitive whose introduction is mandated by Gauthier's explanatory strategy.

IV—Verbal Expressions of Preference Express *Commitments* to Choice Behavior

The thing to do with such a challenge is not simply to contemplate it, but to take it up—to answer the question it asks. I have a hypothesis about what verbal-expressions-of-preference are expressions of, and so how merely attitudinal preferences are related to behavioral ones. Here it is: the distance Gauthier opens up between choice and preference, by acknowledging merely attitudinal preferences as what are expressed by verbal-expressions-of-preference is a *normative* distance. The relation between behavioral and attitudinal preference is the difference between *dispositions* to choose and *commitments* to choose. What verbal-expressions-of-preference express is commitments to patterns of choice behavior.[15]

---

[15] The remarks below are given a much fuller context and more careful exposition in my *Making It Explicit* [Harvard University Press, 1994], especially sections IV-VI of Chapter Four. In particular, the

The idea is that when things go right, when attitudinal and behavioral preferences coincide, the commitments to choose that agents verbally undertake are accompanied by dispositions to make just the choices they have committed themselves to make. In saying I prefer listening to Bach to listening to Brahms, I am committing myself, *ceteris paribus*, to choosing to listen to Bach rather than to Brahms, should such a choice arise. When all goes well, I am disposed to act in the way I have committed myself to acting. On the other hand, I can acknowledge commitments to do things that I am not in fact disposed to do—just as I can promise to do things I am not in fact disposed to do. Attitudinal preferences, construed as commitments to choose, can outrun or even conflict with behavioral ones. Collision of attitudinal and behavioral preferences, of the sort epitomized by the akratic are collisions of a familiar sort: collisions between commitments (or obligations) and dispositions.

So construed—as commitments to patterns of choice behavior—the attitudinal preferences that are expressed by verbal-expressions-of-preference (and so indirectly, those expressions themselves) inherit a structure of entailments and incompatibilities from the outcomes over which those choices are defined. Thus if outcome A entails outcome C, then just as choosing A for that reason entails (will-one, nill-one) choosing C, so *committing* oneself to choose A for that reason entails (will-one, nill-one) *committing* oneself to choose C. And if outcome A is incompatible with outcome B, then just as choosing A is for that reason incompatible with choosing B, so *commitment* to choosing A is for that reason incompatible with *commitment* to choosing B. This is the relation,

---

way I think we ought to understand the relevant sense of 'commitment' is laid out and defended there in a

queried above, between the outcomes over which behavioral choices are defined, on the one hand, and the contents of the words used in verbal-expressions-of-preference, on the other.

What makes a word *mean* 'prefer' (or 'desire', 'want', etc.), on this account, is its use as a primary verb to form declarative sentences whose assertion acknowledges (in the first person case) or attributes (in the second or third person case) a commitment to a pattern of choice behavior. What pattern, exactly? With an eye to the second stage in the process Gauthier envisages, the one at which preferences take the form of *reasons* for action, final or considered preferences, it will turn out to be best to think of the patterns involved in the first instance as patterns of *practical reasoning*. Since endorsement of plans of action, including choice, are the conclusions of bits of practical reasoning, patterns of this sort will entail patterns of choice behavior.

The motivating thought could be put this way. In saying "I prefer A to B,"—that is, in expressing a *prima facie* attitudinal preference for A over B—I commit myself to a pattern of practical reasoning of the form: If doing X is necessary and sufficient for producing outcome A, and doing Y is necessary and sufficient for producing outcome B, and one cannot do both X and Y, then (in the absence of competing commitments) do X. This is a *pattern*, because the commitment I am undertaking or acknowledging by avowing my (attitudinal) preference is indifferent as to what actions X and Y are. It is a pattern of *practical reasoning* because the commitment in question inferentially links doxastic

---

way precluded by the scope of the present essay.

premises (expressed by assertions) to a practical conclusion.[16] That is what is codified above in the conditional inferential structure of antecedent and consequent. Commitments to such patterns of practical reasoning include commitments to patterns of choice behavior, for to reason practically according to any instance of the pattern one commits oneself to by an assertion of the relevant sort *is* to choose one way rather than the other in the circumstances described by the antecedent of the conditional that propositionally codifies the practical inference in question.

One further observation must be registered in order to pick out the patterns of practical reasoning expressing commitment to which is necessary and sufficient for a linguistic locution to qualify as a verbal-expression-of-preference. The patterns of practical reasoning, commitment to which are acknowledged or attributed by first and third person uses of verbal-expressions-of-preference, are *agent specific*. In acknowledging *my* commitment to reason practically (and so to choose) according to the schema of the previous paragraph, I say nothing at all about what anyone *else* is committed to. And if you attribute to me that same commitment, you likewise say nothing about what anyone else is committed to. This fact is the formal reflection of the claim that preference is a subjective matter.

Understanding verbal-expressions-of-preference as whatever has the characteristic expressive role of putting into the form of assertions acknowledgments and attributions of commitments to agent specific patterns of practical reasoning of the structure

[16] I discuss this way of thinking about practical reasoning in more detail in "Action, Norms, and

schematically indicated above—and so understanding attitudinal preferences as the commitments such assertions acknowledge or attribute—answers the challenge put forward at the end of the previous section. It explains the sense in which attitudinal preferences can be decoupled from, outrun, or conflict with, behavioral preferences, as required by a general theory of practical rationality that wants to consider akratic choice (or cyclical preferences) as at least intelligibly describable. And it explains the tight connection between attitudinal and behavioral preference, in virtue of which they deserve to be seen as species of one genus. Thus it allows us to say what must be true of a locution—whether in our own language or in another—for it to mean 'prefers'. Furthermore, as we shall see, the account of verbal-expressions-of-preference as expressing commitments to patterns of practical reasoning and hence of choice behavior meshes nicely with the deliberative process invoked by Gauthier's second distinction among kinds of preference, the distinction between the *raw* preferences (whether merely attitudinal, or also behavioral) agents may just find themselves with, and *final* or *considered* preferences, which alone can serve as genuine *reasons* for action.

V—Preferences as Candidate Reasons for Action

The hypothesis just presented introduces a notion of *commitment* not to be found in Gauthier's text. Even if it is acknowledged that the questions put to Gauthier ("What are

---

Practical Reasoning," forthcoming in James Tomberlin (ed.) *Philosophical Perspectives VII*. [ref.]

verbal-expressions-of-preference?" "How is what is expressed by them related to behaviorally revealed preferences?") are genuine and important, we ought to be sure that the resources he does provide for answering them have been exhausted, before exploring the consequences of supplementing his account in this way. And it may well seem that those explanatory resources have not been exhausted. For Gauthier's second move—introducing the process by which raw preferences are transformed into reasons for action—offers a further, potentially telling, characterization of the role played by attitudinal preferences, beyond that considered thus far. Gauthier needs verbal-expressions-of-preference available to express even behavioral preferences, so that the latter can play a suitable role as inputs in the process of forming *considered* preferences. (And, recall, he needs attitudinal preferences expressed by verbal-expressions-of-preference but not revealed in dispositions to choice behavior, in order to treat as intelligible—albeit irrational—such prime elements of the subject matter of sophisticated theories of practical rationality as the phenomena of weakness of the will and cyclical preferences.)

This observation suggests that the notion of verbal-expressions-of-preference might be explained by appeal not to what lies upstream—that is, by invoking an antecedent understanding of what is avowed or attributed by their use—but by appeal to what lies downstream: the role they play in the formation of *considered* preferences. That role is as potential reasons for action. For, given his account of the process of consideration, nothing could count for Gauthier as a verbal-expression-of-preference (the explicit verbal formulation of a raw preference) unless it also counted as formulating at least a candidate

reason for action—a consideration that, if it is determined to be stable under experience and reflection in the light of other such considerations, provides a genuine reason to act. So when we ask what the merely attitudinal preferences that are only manifested in verbal-expressions-of-preference have in common with behavioral preferences that are also revealed in dispositions to choice behavior, in virtue of which they deserve to be classed as species of a genus recognizable as *preferences*, one available answer is that they are alike in the role they play as inputs into the process of consideration, whose outputs are reasons for action. According to this line of thought, 'raw preference' just means *nondoxastic input to the process of consideration whose product is reasons for action.*

The idea would be to understand preferences just as whatever needs to be added to beliefs to yield potential reasons for action. Davidson, for instance, would insist that the belief that it is raining, even together with the belief that only opening my umbrella will keep me dry, does not yet provide a reason for me to open my umbrella. To have a complete (candidate) reason to do that, we need to supplement those beliefs with a preference for staying dry. The verbal expressions of belief (claims), together with verbal expressions of preference ("I desire, or prefer[17], to stay dry,") together serve as premises in a potentially good piece of practical reasoning whose conclusion is the verbal expression of an intention to act: "So, I shall open my umbrella." (The practical inference is only "potentially" good, because collateral premises in the form of *other* complete reasons for incompatible actions may, upon consideration, turn out to override the preference in question.) The thought

[17] The formal distance between preferences, which are comparative and therefore attach to ordered dyads, and desires, which are categorical and therefore attach to unary objects, can be bridged by understanding a desire to φ as a *ceteris paribus* preference to φ rather than not to φ.

being considered is that one might *define* verbal-expressions-of-preference in terms of the role they play in this process.

This thought is compatible with the suggested construal of attitudinal preferences as commitments to patterns of practical reasoning, and hence to patterns of choice behavior. But neither entails the other. So why not adopt this understanding of raw preferences, in terms of their role in the sort of practical reasoning characteristic of the formation of the considered preferences that provide an agent with genuine reasons for action? The short answer is that this response is not available to Gauthier for two related reasons. First, it gets the order of explanation the wrong way around. Second, the notion it defines is far too broad to be recognizable as a notion of preference at all. A weaker way of putting the second point is that if one could justify characterizing the outputs of the process of consideration as considered *preferences*, rather than just reasons for action, then one would indeed have a warrant for also calling the inputs to the process 'preferences'. But if *all* we know about those outputs is that they are the acknowledgments of the *values* maximization of which is the essence of practical rationality (to which the bulk of Gauthier's story is addressed), then, as we will see below, we cannot distinguish between preference-based approaches to practical rationality and the more committive minimal kantian ones. For in that case: "one might suppose that the quantity to be maximized was not a measure of but a standard for preference."[18]

---

[18] *MBA* 25.

The first point is that adopting the definition of raw preferences in terms of their role as fodder for transformation into reasons for action means that the notion of *reason for action* is not explained in terms of an antecedently specified notion of *preference*. Rather the idea is to christen 'preference' whatever provides a reason for action. That requires a conceptual grip on the notion of reasons for action that is antecedent to our understanding of preferences. Providing an account of this shape is the essence of minimal kantian approaches to action: having a reason for action is acknowledging a value, being rational is choosing (revealing behavioral preferences for) acknowledged values. Refusing to recognize anything as a genuine reason for action for an agent that is not grounded in the preferences of the agent is the essence of purely instrumental, rational choice approaches to action and practical reason. The broadly humean approach to practical reason that Gauthier identifies with is the analogue on the conative side of the empiricist cognitive principle that there are no reasons for belief that do not originate in the senses: *nihil in intellectus est sed fuit prius in sensu*.[19] It is the claim that are no reasons for action that do not originate in inclinations. Minimal kantianism opposes to this a view that makes reasons for action (acknowledgments of value—or obligation) codify, as Gauthier puts it in the passage quoted above, not *measures* of inclination or preference, but *standards* for

[19] The claim that there is nothing in *action* that is not previously in *felt preference* is, like that of its analogue *cognitive empiricism*, crucially ambiguous between a *causal* reading of 'prius' (in the conative case, a *motivational* reading) and a *justificatory* one. One worry is that the pull of conative empiricism as a story about rationality is due precisely to this conflation. Minimal kantianism, as here described, explores a *conative rationalism. Cognitive rationalism* insists that one must already have concepts in order properly to perceive anything. The conative variety claims that one must already have principles and be able to *endorse* them, in order to have inclinations (preferences) in a sense that suits them to serve as *reasons*.

it.[20] Gauthier shares with the more orthodox rational choice tradition commitment to an order of explanation that requires that the concept *reason for action* be explained in terms of the concept of *preference*, which accordingly must be intelligible antecedently. So Gauthier cannot define the species "attitudinal preference", and therefore the genus "preference", in terms of the concept *reason for action*.

The second point is just that there is nothing in this approach to rule out as candidate reasons for action—inputs playing the role of 'raw preferences'—considerations that do not have the right shape or significance to be called 'preferences'. To see this, we need to look a little more closely at the process of forming 'considered preferences'. What is the process of consideration? One important element is assessing the consequences of acting on one's preferences, and of securing the preferred outcome. Assessing those consequences involves inferentially bringing to bear both one's collateral beliefs and one's other preferences, as auxiliary premises in extracting consequences from the outcomes valued in the preferences being assessed. The core element in forming considered preferences from raw ones is inferentially extracting consequences from the rawly-preferred outcomes, on the basis of one's collateral beliefs, and assessing those consequential outcomes on the basis of collateral raw (or, recursively, considered) preferences. I may be inclined (or committed) to choose outcome A over outcome B, but if, in the context of my other beliefs,[21] A entails C, then raw preferences concerning C bear on my endorsement of A. After all, C might be incompatible with the obtaining of the

[20] A more full-blooded kantian view denies that inclinations provide even the raw materials for the practical commitments (endorsements of principles) that are the only genuine reasons for action. This point is discussed below, in connection with the description of intermediate approaches.

even more preferred outcome of some other choice. This is to say that, in the context of an agent's other beliefs, some raw preferences can provide *reasons* either to adopt or to relinquish other preferences. The reasons for action that emerge from the process of considering each candidate reason in the light of others, and in the context of concomitant beliefs, are preferences *endorsed* after consideration of the rest of the agent's attitudes, both cognitive and conative. We need to think of all the preferences involved in this process as propositionally contentful, which is to say *attitudinal* preferences (some of which perhaps corresponding to behavioral preferences that are also revealed by dispositions to choice behavior), manifestable in verbal-expressions-of-preference, because it is a process of practical *reasoning*—and what can play the role of premise in such reasoning (when explicitly codified, the role of antecedent of a conditional) is propositionally contentful, and can be expressed by a declarative sentence.

This process is readily intelligible in terms of the suggested model of attitudinal preferences as commitments to patterns of practical reasoning, and hence choice behavior. Considered preferences will be construed as attitudinal preferences whose endorsement survives competition with other candidate commitments through the exploration of their inferential relations and incompatibilities in the light of beliefs about how things are and how they might evolve under various hypothetical circumstances. But what matters in the present context is that there is nothing about the process of consideration whereby tentative or *prima facie* practical commitments develop into genuine reasons for action that restricts the practical commitments serving as its inputs to those with a structure that

[21] cf. *MBA* 29.

makes them recognizable as preferences (raw or not). For all that has been said so far, inputs to this process could include practical commitments in the form of promises, or obligations incurred because of one's institutional role, or the existence of rules or laws.[22] Nothing rules out acknowledged commitments or obligations of these sorts from operating as considerations in the process of coming to acknowledge reasons for action, even where presystematically we want to say that there are no corresponding preferences: where the agent does not care about fulfilling promises, keeping her job, or obeying the law. Such a situation, of course, is just the one envisaged and endorsed by the minimal kantian.

Why not take a hard stipulative line, and insist that even what are apparently expressions of obligations or commitments, are, when playing this role in practical reasoning, implicitly expressions of preferences? Because their role in practical reasoning is different. "I prefer to stay dry," which genuinely is an expression of preference, at most provides *me* with a reason to open my umbrella (at least absent second-order preferences on the part of others, for instance to prefer doing what I prefer doing). Obligations linked to institutional status, such as that expressed by "Bank employees are obliged to wear neckties at work," by contrast, insofar as they by themselves provide reasons at all (see below) provide reasons equally for anyone who is a bank employee. Other norms, such as that expressed by "It is wrong to (one ought not) amuse oneself by torturing helpless strangers," insofar as they provide reasons at all, do so in a way that swings free of parochial considerations of social status or institutional role. The procrustean strategy of stretching the notion of preference so as to encompass norms of these disparate sorts—

[22] See the discussion of different patterns of practical reasoning in *Making It Explicit*, Chapter Four,

perhaps by distinguishing individual preferences from various sorts of group preferences—relinquishes all continuity with the classical economic notion of behavioral preference, and empties the concept of *preference* of its distinctive content.

Put another way, if we start our theorizing with a bare notion of a nondoxastic input to the process of consideration whose product is reasons for action, the strategy of restricting those inputs to *individual preferences*—which is the essence of the instrumentalist, rational choice paradigm of practical reasoning—will be *ad hoc* and unmotivated. The minimal kantian claims that a good reason for wearing a necktie to work could take the form of the conjunction:

I am a bank employee, going to work.

&

Bank employees are obliged to wear neckties at work.

By contrast, the broadly humean line that traces all reasons for action to individual preferences insists that such a conjunction still falls short of formulating a reason for an agent to act. When fully stated, such a reason must include reference to the preferences or desires of the agent. What more is needed is something like:

I prefer (or desire[23]) to fulfill my obligations as a bank employee.

Section V, and in "Action, Norms, and Practical Reasoning", op. cit..

[23] For a way to fill in the implicit ellipsis, see note 17 above.

Or if the obligation cited in the second conjunct is taken as partly constitutive of the status invoked by both premises, and so implicitly or by implication invokes a possible sanction for noncompliance, perhaps what is wanted is something like:

I prefer to remain a bank employee.

Otherwise, the thought is, the invocation of an obligation consequent upon a social or institutional status does not yet engage with the motivational economy of the agent in the way required to be recognizable as functioning as a reason for that agent.

This is an important, indeed fundamental, dispute—about whether, in what sense, and for exactly what explanatory purposes, internal reasons must be discerned behind external reasons (in one sense of those elastic terms). But it cannot sensibly conducted without some theoretical grip on the notion of preference that is independent of understanding reasons for action—the concept at the heart of the disagreement. For the order of explanation being considered seeks either simply to *define* expressions of preference as whatever is needed to supplement beliefs in order to yield reasons, or to *stipulate* that the reasons in question are to be understood as having to have the form favored by the preference theorist.

VI—Merely Attitudinal Preferences Provide Only Motivationally External Reasons

So it is not open to Gauthier to respond to the question raised here about his first broadening of the classical economic paradigm—including merely attitudinal, as well as behavioral preferences in the raw materials of a theory of practical rationality—by invoking his second broadening—treating raw preferences, whether behavioral or attitudinal, as input to a process of consideration of preferences, whose outputs alone are *prima facie* reasons for action. The issue of how to understand attitudinal preferences—some of which are expressed only in dispositions to verbal behavior, not also in dispositions to nonverbal choice behavior—is not settled by looking at the process of consideration by which preferences are turned into reasons.

The discussion of that issue pointed to an important motivation for preference-based theories of practical reasoning: the thought that whatever we treat as fundamental reasons for action should play a certain kind of role in the behavioral economy of the agent. In particular, they should be intrinsically motivating; what reasons for action an agent possesses should make a difference to what the agent is disposed or inclined to go on to do. This is one of the main points of contrast between broadly humean (empiricist) and broadly kantian (rationalist) approaches to practical reasoning: the kantian insists that something say, a commitment or obligation, can genuinely provide an agent with a reason for action even though the agent is not in the least inclined to act according to it. It is

worth noticing that the minimal kantian need not follow Kant himself in denying that an agent's behavioral preferences, dispositions, or inclinations—Kant's "sinnliche Neigungen"—can be the basis of any genuine *reasons* for action. Where the minimal kantian differs from the humean is in insisting that such inclinations do not *exhaust* the reasons for action an agent may have.

If we draw the lines in this way, however, Gauthier's first move already marks his departure from the humean camp, and his enlistment with the minimal kantians. For that move consists in allowing the possibility that merely attitudinal preferences, not accompanied by corresponding behavioral preferences, can provide (the raw material for) genuine reasons for action. Preferences of this sort are distinguished precisely by not engaging with the agent's behavioral economy so as to be intrinsically motivating. Attitudinal preferences may line up with dispositions to nonverbal choice behavior, but then again, they may not. When they do not, they function as motivationally external reasons. Calling them 'preferences' and pointing to their subjectivity (in the sense that distinguishes them from transpersonal obligations such as those associated with institutional roles such as being a bank employee) does not alter that fact.

It should not be thought that a concession to the minimal kantian along these lines is an optional or idiosyncratic feature of Gauthier's understanding of preference. *Any* account of preference or practical reasoning that can acknowledge *akrasia* as an intelligible form of practical irrationality must make this concession concerning motivationally external reasons for action. The possibility of decoupling what an agent has reason to do (indeed, *acknowledges* having reason to do) from what the agent in fact is disposed to do is the

very essence of the phenomenon of weakness of the will. Of course the decoupling is not complete: when things go right, attitudinal and behavioral preferences track each other. But the decoupling is not complete for those who would include motivationally external norms as first-class reasons for action either. For such theorists would insist that to be a properly trained agent is to be disposed, *ceteris paribus*, to respond to the acknowledgment of a commitment or obligation by fulfilling it. Gauthier is with the minimal kantian both in claiming that genuine reasons for action are not restricted to motivationally internal ones, and that those surplus external reasons need not for that reason be understood as motivationally inert. Both explore the important middle ground between requiring reasons for action to be *intrinsically* (because definitionally) motivating, and being forced to see reasons that fail of that status as therefore motivationally *inert*.

Thus consideration of the dispute about whether motivationally external reasons ought to be included along with motivationally internal ones as basic inputs to the process of practical reasoning underscores the urgency of answering the basic question raised by Gauthier's first move: how should we understand the relation between behavioral and attitudinal preferences, and hence the genus *preference* itself? With the formulation of the previous paragraph in mind, we may ask how we should understand the notions of 'tracking' and 'going right', such that when things go right, attitudinal and behavioral preferences track each other? We may grant that allowing room for slippage—besides being true to the phenomena—need not rule out linkage. But what sort of linkage, and

how should we understand the difference between things going right and them going wrong?

My suggestion, recall, is that this relationship is a *normative* one—just as talk of "things going *right* (as they *ought*)" suggests. An attitudinal preference is the acknowledgment of a *commitment* to act (choose) in a certain way (according to a specifiable pattern). Attitudinal preferences 'track' or correspond properly to behavioral preferences when those attitudes are explicit expressions of the behavioral preferences one implicitly commits oneself to by being disposed to act (choose) in those ways. Behavioral preferences 'track' or correspond properly to attitudinal preferences when one is in fact disposed to act (choose) as one acknowledges oneself to be committed to do. The sort of propriety involved—what is invoked by talk of "things going right"—is in the one direction that of acknowledging one's implicit commitments, and in the other direction that of fulfilling commitments or obligations one explicitly acknowledges.

Attitudinal preferences without corresponding behavioral preferences are commitments to patterns of choice behavior in which one is not in fact disposed to engage—a situation intelligible, indeed familiar, to us from the otherwise very different case of promises, which can coherently (if culpably) be undertaken even where the promiser is not disposed to fulfill the commitment made (in that case) to another. Behavioral preferences without corresponding attitudinal preferences are behavioral dispositions of which the agent is not discursively aware, cannot make explicit in the form of assertible (hence thinkable) expressions-of-preference. Since only such verbalizable expressions-of-preference can serve as premises in practical reasoning, and so are available as inputs to the process of

consideration that will yield genuine reasons for action, such *merely* behavioral preferences are *rationally* inert.

VII—Considered Preferences are *Endorsed* Preferences

Gauthier's second innovation is to distinguish *considered* preferences from *raw* preferences (whether the latter are behavioral or attitudinal). I think that looking at the process of consideration, whereby raw preferences are transformed into genuine reasons for action, offers some confirmation of the normative reading of the relation between attitudinal and behavioral raw preferences. For I think that process should be understood as one in which raw preferences acquire a further sort of *normative* force. The force of *reasons* for action is a matter of considerations serving as premises in practical inferences that can *justify* or *entitle* the agent to various practical commitments (commitments to act).

Gauthier understands the inputs to the process as attitudinal preferences (some of which do, and some of which perhaps do not correspond to behavioral preferences). For only these verbally expressible preferences are explicit in the form of claims, and hence available to serve as *premises* in bits of practical reasoning. Accordingly, only what is in this form can be understood as a candidate *reason* for action—the status acquired by preferences that survive the process of consideration. Why might some of them not survive? Gauthier's discussion of considered preferences is subtle, and its nuances deserve

more discussion than they can be given here. For our purposes it will suffice to consider two of the major dimensions along which preferences can, upon consideration, be found wanting, and hence relegated to the status of not providing reasons for action. First, preferences which would provide reasons for action in isolation, may fail to do so if they occur as part of a set that includes incompatible preferences. Second, preferences that would provide reasons for action if they were stable and persistent, may fail to do so if upon consideration they are deemed to be transitory or short-lived, likely to be replaced, revised, or succeeded by contrary preferences—paradigmatically upon the receipt of further information. These two major ways in which raw preferences may fail to qualify as considered preferences plays an important role downstream in Gauthier's subsequent argument for rationality as constrained maximization of the value that is a measure of considered preference. But each is independently motivated upstream by intuitions and considerations concerning the requirements of practical rationality.[24]

If we restrict ourselves to purely behavioral preferences, incompatibilities cannot arise. For those preferences are implicit in actual choices (or dispositions to make such choices), and those choices resolve any potential conflicts. But the addition of attitudinal preferences makes conflicts possible; indeed, this conceptual broadening is introduced in part in order to make intelligible such collisions of preference as those exhibited by the akratic or the avower of cyclic preferences. What is irrational about agents who find themselves in these situations is that the reasons for action that would otherwise be provided by one preference are undercut by the presence of contrary preferences. Until and unless the conflict is resolved, the agent cannot be said to be acting for reasons at all.

[24] *MBA* 33-38.

This situation makes perfect sense if, as I have suggested, we understand what such verbal expressions express as *commitments* to patterns of action (including choice behavior). I may undertake incompatible commitments to choice behavior, just as I may make incompatible promises—ones that cannot be jointly fulfilled. The akratic undertakes incompatible commitments: the commitments implicit in the pattern of choice behavior he is disposed to are incompatible with others she endorses verbally. The subject who finds herself with cyclic preferences also undertakes incompatible commitments. We want to say that these are two forms of irrationality. What is irrational about them is precisely that the commitments undertaken by the akratic and by the subject of cyclic preferences are jointly incompatible.

Two claims are incompatible just in case *commitment* to one precludes *entitlement* to the other. One is not precluded from undertaking the other commitment (making the other claim); there is nothing incoherent about undertaking incompatible commitments (think of promises). To say that a set of commitments is jointly incompatible is to say that one cannot be *entitled* to all of them.[25] Promising to take you to the airport tomorrow morning at ten does not make unintelligible or impossible my promising to take my son to his crew practice tomorrow morning at ten. But the distance between the airport and the river (along with other uncontroversial background facts) means that the one promise keeps me from being *entitled* to make the other promise. Practical reasoning is about *justification*, about the agent's *entitlement* to various commitments, and so choices or

[25] I discuss this way of thinking about incompatibility in Chapter Three of *Making It Explicit* (op.cit.).

courses of action.[26] Reasons for action are practical commitments to which one is entitled, and which can accordingly *justify* what one does. According to this way of thinking, the process of consideration is the process by which some of the raw preferences are *endorsed*, in the sense of taken as potentially justifying action (including choice). Endorsing incompatible commitments is structurally self-defeating. Just so, on Gauthier's view one's *considered* preferences cannot be mutually incompatible.[27]

The requirement for stability of raw preference, in particular under increases in information, is also intelligible in these terms. The practical commitments expressed by verbal expressions-of-preference (desires) are like the doxastic commitments expressed by assertions (beliefs), and unlike the practical commitments (*to* someone) undertaken by promises, in that one is permitted to change one's mind—to alter or relinquish the commitment. Nonetheless, to take it that the candidate for endorsement (upon consideration) is a *commitment* involves taking what one is committed to by it to include a certain sort of stability over time and circumstance. What is expressed by a statement such as "I believe that neither the word 'problem' nor the word 'solution' occurs in the King James edition of the Bible, but if I looked into the matter further, I would probably

[26] Of course it is also about consequential commitments, for instance how commitment to securing an end can commit the agent to a necessary means. But such committive consequences bring along relations concerning entitlements. For if commitment to bringing it about that *p* entails commitment to bringing it about that *q*, then one cannot be entitled to the first commitment if one is not entitled to the second.

[27] One might object that satisficing considerations could make it rational not to resolve some residual incompatibilities, since the cost of doing so (given one's other preferences) is too high. But in fact, this is need not be seen as an objection. Gauthier's aim is to say what we ultimately have reason to do. When rationality as satisficing dictates that we not resolve incompatible preferences, the situation might be understood as one where economic efficiency means it is not worth putting more effort into further refinements of one's idea of what one ultimately has reason to do—the current take is good enough for practical purposes. (Of course, taking this line would involve considerable work to entitle the theorist to the assumptions about approximation on which it relies.)

change my mind,"[28] is not a *belief*. For the attitude being evinced falls short of taking the contained claim to be true, of *committing* oneself to its truth. One is not treating the claim about those words' nonoccurrence as a premise whose inferential consequences ought to be endorsed. Just so with preferences. What is expressed by a statement such as "I prefer electing the city council by districts to doing so at large, but if I looked into the matter further, I would probably change my mind," is not the endorsement of a preference. One is evincing one's offhand disposition to *say* something, but precisely with the proviso that one is not really endorsing that disposition. One is not putting that disposition forward as a *reason*, in the sense of a premise for practical reasoning whose consequences ought to be endorsed in the sense of enacted.

VIII—Conclusion

Gauthier does not talk this way about the process of turning raw preferences into considered preferences. He has good reasons to resist and reject the proffered transposition of his view into this normative idiom. For one thing, so translated, Gauthier's position would assume very much the shape considered—and rejected on his behalf—in Section V above. For if considered preferences are commitments to patterns of practical reasoning that the agent has endorsed as ones to which she is entitled, given her collateral commitments (both doxastic and practical-preferential, and including those

[28] Those words do not in fact occur in that work. (I believe Raymond Williams was the first to point out

concerning the likely evolution of her dispositions to undertake such commitments under the influence of extensions of her knowledge, assessment of the consequences of acting on those commitments, and so on), then it is hard to see a principled reason to restrict the inputs to the process of consideration to commitments that have the structure characteristic of individual preferences. In deciding which considerations to endorse as having the force of reasons for action, why should the agent not consider also commitments that do not have the shape of preferences, such as promises made, or obligations incumbent on social or institutional status—even when not accompanied by corresponding preferences? Although there is often confusion on this point, the fact that *if* the agent had *contrary* preferences (for instance, positively *wanted* to be fired from her job at the bank) these might well override acknowledged commitments, does not entail that the commitments provide reasons for action only when combined with preferences for fulfilling those commitments. Nothing in this way of thinking about things entails Kant's own view, that merely behavioral preferences are in principle not eligible for endorsement as reasons.[29] But if what one is doing in considering raw preferences is deciding which commitments to endorse, what rationale is there for restricting that consideration to commitments that have the special structure of individual preferences?

Of course the only commitments that matter to the practical reasoning of an agent are commitments she *acknowledges*, or treats as in some sense binding on her. But to say that

---

this significant difference between our habits of mind and those of its time.)

[29] There is an asymmetry in Kant, in that he allows that we can act morally, apart from all reference to inclination, but not that we can think apart from all reference to intuition. The view I am calling 'minimal kantianism' allows for intuition and inclination to play more parallel roles. I think Hegel has a view like this, which envisages a process of *conceptualizing inclinations* parallel to that of

is not to say that they must be commitments she *prefers* to keep. For a preference is a commitment that binds only the one whose preference it is. In acknowledging that bank employees have an obligation to wear neckties at work, one is acknowledging an obligation that is part of playing a certain institutional role. What one acknowledges is something binding on anyone having that status, *as* something binding on anyone having that status. Similarly one might acknowledge an obligation or commitment not to amuse oneself by torturing helpless strangers as binding on agents in general. In doing that, one is acknowledging something whose bindingness is not taken to depend on anyone's acknowledgment of it—something that is not a matter of preference. Nothing in Gauthier's discussion of the process of consideration by which raw preferences acquire the status of reasons for action precludes, or even motivates precluding, other sorts of considerations besides *preferences* from serving as inputs to this process. His account is one the minimal kantian can endorse, adopt, and applaud.

In *Morals By Agreement*, Gauthier sees himself as continuing the broadly humean, empiricist tradition of classical economic theories, which understand rational choice in terms of maximization of utility, the measure of individual preference. He explicitly aligns himself with this camp, against those who would understand some source of reasons for action beyond the preferences of individual agents. But each of the dimensions along which he finds that the phenomena of practical reasoning oblige him at the outset of his enterprise to broaden the classical understanding of preference, I have argued, is best understood as implicitly introducing normative notions that move substantially beyond that

---

conceptualizing intuitions. These are mutually involving aspects of the other side of the process of *sensualizing concepts*, by which alone they acquire determinate content. But this is Hegel, not Kant.

picture. In fact each of his basic moves—recognizing attitudinal, and not just behavioral preferences, and recognizing only considered, and not raw preferences, as constituting reasons for action—opens up the space it is characteristic of minimal kantianism to insist upon. Attitudinal preferences are what are expressed by verbal expressions-of-preference. I have argued that Gauthier does not tell us what they are. But I have suggested further that we should understand them as *commitments* to patterns of action or choice behavior. Considered preferences, further, should be understood as such commitments that are *endorsed* as *justifying* or *entitling* the agent to act.

Preferences constrain how action *ought* (in a distinctive sense) to go. The distance between behavioral and attitudinal preferences (so important for understanding the possibility of the distinctive form of irrationality that is weakness of the will) is a normative distance: the difference between dispositions or inclinations to act or choose, and acknowledged commitments to do so. Acting according to this sort of commitment (made explicit by an 'ought') is being rational (having a rational will). But to see the distance that makes acting this way a task at which one could fail requires introducing the notion of a normative status, a kind of *commitment*, that is not intelligible in terms of the economist's—nor indeed, of Gauthier's—notion of maximization. For it is *presupposed* by any notion of maximizing that starts with preferences understood in Gauthier's extended sense. Again, the notion of *endorsement*, and the corresponding sense of *entitlement* to a course of action that can be inherited from a commitment, which is what the notion of a reason for action comes to, is not one that can be understood in terms of

maximization. For it, too, is part of the raw materials—the understanding of preference and value—in terms of which maximizing can then be introduced.

In the years since *Morals By Agreement* appeared, Gauthier has in many ways moved in a kantian direction. I think that if we press hard enough on the question of my title—if we ask what it is that is expressed by verbal expressions-of-preference—it becomes open to us to see that this development not so much as an alteration of his views, as the becoming more explicit of what was already implicit in the notion of preference with which he began his enterprise. But that possibility raises a final reason for the author of *Morals By Agreement* to reject the normative answer offered on his behalf to the question of my title, an answer which pushes his view substantially in the direction of minimal kantianism. For in this essay I have (possibly perversely) completely ignored the crucial third stage of Gauthier's story: I have said nothing whatsoever about constrained maximization. What we *finally* have reason to do, according his story, is not what is recommended by our considered preferences, but only what emerges from constrained maximization of those preferences. Put another way, Gauthier offers *three* stages in the grooming of reasons (raw preferences, considered preferences, and what maximizes considered preferences under the constraints it is his main task to argue for), while the account reconstructed here offers only *two* (commitments to act, and entitlements to those commitments). In effect, the minimal kantianism reconstructed here out of the raw materials Gauthier provides collapses the final two levels that he is concerned to distinguish. So the line of thought presented here, which begins with the question posed in the title of the essay, ends by posing a further question, which cannot be pursued here: How might the minimal kantian

for which I have been arguing Gauthier makes room understand the distinction between two sorts of entitlement to practical commitments, which he urges on us in distinguishing considered preferences from final reasons for action?

Bob Brandom

University of Pittsburgh