Constructing the Inner Citadel: Recent Work on the Concept of Autonomy

John Christman

The metaphor of an "inner citadel" was not used approvingly when Isaiah Berlin invoked it as part of his attack on the notion of positive liberty (Berlin 1969, pp. 135 ff.). And although he scoffed at the idea that there loomed within us all an inner structure representing our "true" selves, he was moved to admit that the idea of self-government that positive liberty is meant to capture is a value to be counted among the various elements of a free society. After all, can a person be said to be free or acting freely if the desires which produce those actions do not bear the pedigree of authenticity—the person's "true desires"? What account can we give of the self-critical and self-mastering capacities of a person utilized to form the desires which motivate free action? What such an account would amount to, then, is a theory of individual autonomy. And while a great deal of attention has been paid recently to this notion, much of the work done on the ideas of self-government and autonomy avoids the use of the term 'autonomy.' Nevertheless, what I wish to do in this survey is to touch on the various attempts at constructing such theories of autonomy (or what amounts to autonomy), and the resulting controversies that these have spawned. After asking whether it makes sense to talk of a single or central notion of individual autonomy at all, I will discuss some of the most influential recent theories of autonomy and the criticisms that these have faced. In Section II, I will touch on the relation of autonomy to other values (in particular utility and rights) and turn finally to the question of the value of autonomy itself.

I. THE CONCEPT OF INDIVIDUAL AUTONOMY

Is There a Unifying Idea?

Now as I admitted, some of the writers discussed here avoid the use of the word 'autonomy' altogether in their work, but there are various, seemingly unrelated problems that are connected in that they all revolve around this core notion of individual autonomy. This is not to presuppose an obviously controversial claim that 'individual autonomy' has a single meaning that everyone who uses the term is aware of. It is only to focus in on the notion of self-government that underlies at least the central use of the concept.

Feinberg (1986, chap. 18) is explicit in his doubts that 'autonomy' has a single, coherent meaning. In this light, he attempts to explicate the related notions that function to formulate the general conception of "personal autonomy." He claims that "the word 'autonomy' has four closely related meanings" which refer either to the "capacity" to govern oneself, the "actual condition" of self-

Ethics 99 (October 1988): 109–124
 0014-1704/89/9901-0010$01.00

government, an ideal of virtue derived from that conception, or the "sovereign authority" to govern oneself. A full theory of autonomy, he suggests, would spell out the relations among these different meanings of the term and presumably support the various implications of the notion of autonomy in its different guises. He then proceeds, with characteristic thoroughness, to illuminate the sundry "virtues" that exemplify the condition of autonomy. The list includes qualities such as moral authenticity, self-legislation, distinct self-identity (individuality), and self-control (pp. 28 ff.).

Feinberg is not explicitly averse to the contention that the traits of "authenticity" and "self-determination" he discusses might straightforwardly be read as the core ideas of which the other characteristics in his list are extensions and variations (pp. 28 ff.). Indeed, the four "meanings" of autonomy listed by Feinberg all center on just such a conceptual core: the actual condition of autonomy defined as a psychological ability to be self-governing. If I am trying to give an account of some property X, I will be looking for the major conditions of X's actually obtaining. It is not to deny this to point out, as Feinberg does, that there are related concepts corresponding to: a *capacity* for X; seeing X as a *character ideal;* and the supposed *right* to X. These are ideas whose meaning must rest on the central account of X itself.

There is of course a wide variety of uses of the concept of individual autonomy which are extensions of or related to the idea of an actual condition of (something that amounts to) self-government. The latter notion is a descriptive property instantiated by some or most human beings. However, as Hill points out (1987, pp. 133 ff.), 'autonomy' is used sometimes to pick out, not the actual psychological condition of self-government (PC-autonomy), but rather a *right* not to be treated in certain ways. Specifically, autonomy as right (R-autonomy) is a right against actions which attempt to disrupt or undercut one's PC-autonomy. When a person brainwashes me, that person violates my R-autonomy by interfering with my ability to critically evaluate my desires and choices. Such is also the case with many threats, manipulations, and acts of violence. These acts interfere with my ability to control a certain area of my own life that should be left strictly to me.

There is an additional ambiguity, however, in the analysis of cases where a person's R-autonomy is violated. In some instances, my R-autonomy is violated even when the capacity to critically evaluate my own choices is not actually disrupted; it is violated in cases where I have been treated *as if* I had no PC-autonomy or without sufficient *respect* for my PC-autonomy. Consider the way parents might violate the autonomy of, say, their adolescent child by constantly preempting her choices about clothes, a place to live, what car to buy, and the like, and buying these things for her secretly and surprising her with them. The parents violate the person's R-autonomy not because they prevented her from freely analyzing and identifying with her desires and tastes (i.e., blocking her PC-autonomy) but, rather, by treating her as if she could not adequately do so herself.

Meyer (1987) draws a somewhat similar distinction in his discussion of "negative" and "positive" conceptions of autonomy. On a negative view, he suggests, a person is autonomous when she is not directed in some crucial way by another person. One is autonomous in the positive sense when one is actively self-directed. I suspect, however, that these two notions ultimately collapse. A full specification of what it means to be *self*-directed, in a manner that captures what it means to be autonomous, simply will include the sorts of factors (or the conditions for such factors) that must be *absent* for such self-direction to occur.

Another distinction Meyer makes, however, comes closer to the difference between R-autonomy and PC-autonomy. Meyer refers to two positions on the nature of autonomy, one he calls the "Stoic" position, which asserts that autonomy is only maintained when the person secures active control over her own (lesser, lower?) urges and impulses (p. 267). On the other hand, "Rights Sensitive autonomy" is a normative property of persons whose liberty to choose their own way of life is protected (pp. 267–68). While this latter notion is quite similar to R-autonomy, the "Stoic" position remains somewhat obscure. Stoics could be claiming one of two things: one (similar to Kant) is the claim that autonomy obtains only when a person is not guided by the unwanted urges and desires she finds herself with (or like Kant, by desires at all); or the position is that one can maintain autonomy in the face of obstacles to action by reshaping one's preferences to better conform to the set of feasible options, thereby achieving the highest ratio of satisfiable to nonsatisfiable desires. But this latter view, which is closest to the view of the actual Stoics, cannot by itself determine the autonomy of these preference changes. For what determines whether a person is autonomous in the revision of her set of preferences depends crucially on *why* this change comes about. As Elster argues (1982), there is all the difference in the world between mere "sour grapes" (altering desires because you can't get what you want), and conscious character planning (deciding to mold one's character to meet limited possibilities), the latter being consistent with autonomy. But the Stoic strategy, by itself, tells us nothing about why the person is adopting the strategy (was it forced upon her by a relentless Epictetian brainwasher?), and this *external* question is the critical determinant of the autonomy of such decisions.

Another question about the structure of the property that autonomy is meant to pick out concerns its scope. One way of construing autonomy is that, at its most basic level of application, it is properly seen as a property of preferences or desires (or their formation),[1] while another view is that it is a property of whole persons or of persons' whole lives. This first approach effectively connects with the debate over the "endogeneity of preferences" found in the economics and social science literature (cf., e.g., Braybrooke 1974; Sen 1977; Yaari 1977; and Elster 1982, 1983). Others of course cling to the idea that autonomy is a property of whole persons or persons' lives (Dworkin calls it a "global" concept), and hence it cannot be applied at any more localized level (cf. Dworkin 1981; and Young 1980, 1986).

Also, the relation between autonomy and freedom (of action) is one that often goes unexplicated (or the notions of freedom and autonomy are simply used interchangeably). When freedom is construed as primarily a property of human *action,* the account typically assumed is that freedom is the relation among the agent, desired actions, and some set of restraints.[2] But no matter how rich a conception of "restraint" one works out in this context (that one must be "free from" to be free when one acts), it will always be a further question whether the desire a person is acting on is autonomous or not. (For more on this point, cf. Neely 1974 and Arneson 1985.) A person acting on the basis of a nonautonomous preference (placed into her brain through hypnosis, say) faces no restraint in

1. Throughout this discussion, I use the word 'preference' very broadly to refer to desires, values, or "pro" attitudes an agent might have.

2. The "triadic" conception of freedom is discussed by MacCallum (1967), Feinberg (1973), and others.

performing the action which is the object of the preference. In fact, the language of restraints is wholly inappropriate here, because the person is not being stopped from doing the act, she is being *forced* to do it. So although the term 'autonomy' is sometimes used in this wider sense (freedom of action), a (perhaps) more basic notion of autonomy will give an account of how a person came to be the way she is in determining her freedom (in the fullest sense). A full formula for freedom (of action) then might be proposed: to be free (in a given context) means there is an absence of restraints (positive or negative, internal or external)[3] standing between a person and the carrying out of that person's *autonomously formed* desires.[4] While this claim needs more defense, it can at least be suggested that many of the disagreements I've cited here indicate the need to ask whether a coherent meaning can be set out for this core idea of self-government which these views revolve around.

An Account of Autonomy

Much work has been done in recent years to try to set out the conditions of individual autonomy coincident with intuitions about self-government and freedom. The work of Dworkin (1976, 1981) and of Frankfurt (1971) are seminal in this regard. More recently, Young (1986), Lindley (1986), and Haworth (1986) have produced book-length treatments of the concept. I will begin with a discussion of the Dworkin/Frankfurt model (the DF model), and indicate the refinements introduced by Young, Lindley, and Haworth.

The DF model rests on the notions of lower-order and higher-order (or first-order, second-order, etc.) desires. The theory of the person worked out in Frankfurt's paper and the explicit theory of autonomy put forth by Dworkin build on this distinction. Lower-order desires (LODs) have as their object actual actions of the agent: a desire to *do* X or Y; higher-order desires (HODs), however, have as their object other, lower-order desires: a desire to desire to do X or Y. Some of those lower-order desires, goals, and preferences are ones that an agent will "identify" with in an important way. Building on this idea, Dworkin's "full formula for autonomy" is spelled out this way: "A person is autonomous if he identifies with his desires, goals, and values, and such identification is not influenced in ways which make the process of identification in some way alien to the individual. Spelling out the conditions of procedural independence involves distinguishing those ways of influencing people's reflective and critical faculties which subvert them from those which promote and improve them" (Dworkin 1981, p. 212).

Identification takes place when an agent reflects critically on a desire and, at the higher level, approves of having the desire.[5] Whether such critical reflection must be rational is a question that I will return to below. For now, let us say merely that the authenticity condition is met when the agent accepts the desire, value, or preference as part of her larger set of desires, beliefs, and principles, whether or not this is done for good reasons.

3. These are the four types of restraints mentioned by Feinberg in his analysis of freedom; see Feinberg (1973), chap. 1.

4. This construal of autonomy suggests a direct link to the question of free will and determinism, an issue not surveyed here. For a recent overview of that literature, see Watson (1987).

5. In Frankfurt's case, the requisite higher-order desire is that the lower-order desire moves the agent effectively to act, i.e., it is a higher-order "volition"; see Frankfurt (1971).

Some of the most serious reservations about this model have come from skepticism concerning the vagueness of the notion of identification (cf. Thalberg 1978, p. 220; and Watson[6] 1975). As a response to this, Young has attempted to improve on this condition by utilizing the Freudian imagery of "identification with another person" as a guide to the nature of this component of autonomy. For Freud the paradigm of identification with another person (or another "object," like an ego-ideal) can be described as empathetic imagination of the other person "from the inside" (Young 1986, pp. 43–47). However, a possible problem with following the guiding influence of psychoanalysis in this regard is that the Freudian act of identification can take place completely at the unconscious level, and hence amounts to a process at least as mysterious as conscious identification with parts of oneself. What may help here is what, in a different context, Strawson has called "Integration." One could say (adapting Strawson's view) that an agent identifies with a desire when, from his own point of view, its "being involved in the determination of the action (citable in true rational explanations of it) *just is* his being so involved" (Strawson 1986, p. 245). This amounts to simply acknowledging the desire as *me,* without any corresponding endorsement of it.

But this characterization points to what may be the deeper problem with the condition of identification. Either one sees identification as simple acknowledgement of what desires I find myself with, or one builds into the notion an evaluation of the having of that desire. On the former view, identification can appear to conflict with (an intuitive sense of) autonomy, for I can acknowledge a desire as my own, in a straightforward sense, even though it is not the result of autonomous processes of preference formation. I'm stuck with it, so to speak. But this implies that I judge it to be *me.* If, on the other hand, one strengthens the requirement of identification to include an *endorsement* of the desire, then this would (implausibly perhaps) rule out the possibility of having an autonomous desire I don't approve of. To be autonomous in this way, I would have to be perfect (in my own eyes).

The model we have been considering, however, faces other sorts of difficulties which are discussed by several authors (cf. esp. Thalberg 1978, 1979; and Friedman 1986). One of the major objections is this: we can imagine a person who lives a completely subservient life, and who also identifies with the first-order desires that constitute such a life. Socialization, education, and conditioning throughout the person's upbringing lead her to see, let us say, the life of a subservient housewife as her true calling. Thus, on the hierarchical analysis, she passes the test of autonomy since her HODs are consistent with her LODs. She approves of the LODs and identifies with them. Obviously, however, she is a manipulated individual whose choice of life-style and values are not her own in a real sense. Her values, even at the second level, are the product of her upbringing and conditioning. The most poignant form that this criticism takes is in Thalberg's charge that the requirement of Procedural Independence merely introduces an infinite regress (1978, pp. 219–20). Since the acts of identification must themselves be autonomous, this requires that another act of identification take place at a higher level. And since this act must also be carried out in a way that reflects

6. Frankfurt and Watson are ultimately concerned with the issue of free will in the papers cited, a debate avoided here. However, as I explain in the text, their views can be applied directly to the development of a plausible concept of autonomy.

procedural independence, then a fourth level must be postulated there. Hence the regress.

One needs, of course, to set *some* conditions on the higher-order approval of desires, for it would be implausible to allow that a person can have autonomous first-order desires despite having nonautonomous higher-order desires. If not, the view faces what could be labeled the "ab initio problem," in that it involves the claim that desires can be autonomous without foundations. But certainly a person cannot be autonomous at a lower level of desire when her endorsement of them is the result of manipulation further up the hierarchy of preferences. (For further discussion of this point, cf. Friedman 1986, pp. 22–23; and Christman 1987.)

In a recent article Frankfurt (1987) has attempted to respond to the regress argument by claiming that there is no need for a higher-order endorsement of the initial act of identification (with a LOD) as long as that endorsement is made decisively. "[An endorsement] is decisive if and only if it is made without reservation . . . [that is, it is made] in the belief that no further accurate inquiry would require [the person] to change his mind" (p. 37). Frankfurt goes on to explain and defend this view, but it nevertheless does not look promising as a solution. For no matter how decisive and final a person's identification with a desire is, that identification could still be the result of obviously heteronomous manipulations (hypnosis, say, which included the directive to maintain identification with the desire even when informed of the hypnosis). Hence either we declare the person to be lacking autonomy despite the decisive identification with the desire (in which case identification is insufficient and perhaps unnecessary for autonomy) or we declare such obviously manipulated individuals autonomous (which is subject to what I labeled the ab initio problem). Now Frankfurt could mean that it must be *objectively reasonable* for the person to identify with the desire for her to be autonomous in regard to it, which raises the whole issue of rationality as a requirement for autonomy. This is a question I will return to below.

Another problem with the DF model, one raised by Friedman (1986), is the claim seemingly implied by the conditions of the model, that higher-order desires make up the "true self." But why should this be true, Friedman asks. Imagine a person whose higher-order desires (her "principles") have been conditioned to conform to some oppressive standard of behavior. But imagine also that the person still has first-order impulses that resist these "principles" (she hates to do the dishes, say, but forces herself to dutifully do them out of regard for her conditioned role). This shows that the higher-order preferences of the person actually reflect much *less* the nature of her "true self," in some intuitively compelling way. Now a response to all these problems (the regress problem, the ab initio problem, and the "true self" problem) could be the spelling out of some further condition bearing on the act of identification which is not simply a higher-order approval, but nevertheless captures the idea of self-government. Whether this can be done to anyone's satisfaction is still an open question, but such an account need not necessarily attribute special metaphysical status to any of the "selves" involved in the process of self-evaluation. (For more on this suggestion, cf. Christman 1987.)

The Kantian Tradition: Autonomy and Rationality

Whatever contemporary approach to the topic of autonomy one finds appealing, the debate over the concept will inevitably be framed by the baggage inherited from Kant's theory of autonomy. On Kant's view, one is autonomous if one is in a position to subject one's will to self-imposed maxims which conform to the

moral law. That is, Kant saw the relation between autonomy and laws of reason to be a particularly intimate one. While the generic idea of "self-rule" is still essential to recent views of autonomy, contemporary theorists treat it as an open question whether agents can be autonomous if they fail, vis-à-vis the act or desire in question, to follow the dictates of morality or reason. Indeed there are many aspects of Kantian autonomy that have been abandoned by recent philosophers, as Hill (1984) points out. For example, writers no longer see autonomy as a property of the "will" per se, where that functions independently of desire. Second, philosophers of late have tended to avoid Kant's view that autonomy is a property (of the will) which lies outside the causal nexus of space and time (it occurs at the level of the noumena), and hence not subject to empirical explanation. Most remain neutral on the question of the metaphysical status of the processes of free choice (except, that is, those writers who specifically address the question of the relation between freedom and determinism).

As I indicated, though, what has divided many philosophers in relation to the Kantian conception of autonomy has to do with the connection between autonomy and reason or rationality. Haworth (1986), for example, has developed a model of autonomy that relies heavily on the reasoning capacity of agents in sifting through their own desires. In Haworth's account of "normal autonomy," a trait developed naturally by a typical adult, he includes the condition of "critical competence" that an agent must exhibit to be autonomous. Critical competence, on this view, demands that one go "far enough in finding reasons for one's preferences, [but] without needing to go to heroic lengths of deliberating endlessly" (p. 39). An autonomous person must display a degree of what Haworth calls "full rationality." This includes more than "means-ends" rationality, but also the ability to subject one's ends themselves to critical appraisal in light of one's other beliefs and values, future preferences one expects to have, and present desires one has about the future. This model, then, puts less weight on the requirements of Procedural Independence (though that is a condition) than on the demand that an autonomous person be a minimally good reasoner.

Benson (1983) articulated this approach most forcefully in his contention that "intellectual" autonomy demands that one "put oneself in the best position to answer for the reliability of one's beliefs" (p. 8). Similarly, Meyers (1987a) develops what she labels a "competency" approach to autonomy. On her view, "personal autonomy . . . can only be achieved through the exercise of a repertory of coordinated skills" (p. 627) which are necessary to be competent in self-reflection and criticism. In this same vein, Lindley (1986) discusses the contrasts between minimal, Humean conditions for rationality and more stringent Kantian requirements. Although on his view neither, as it stands, is a plausible requirement for autonomy. He objects to the Kantian claim that to be autonomous is to be rational in the "pure sense," which is to be motivated by reasons untainted by desire or inclination and which are not "situation specific" (i.e., they are universalizable). For Lindley, this strong sense of rationality is too stringent a requirement for autonomy for, as he claims, "autonomy is primarily a matter of authorship, [while] rationality is essentially a matter of acceptability" (p. 21). The Humean view of rationality, on the other hand, is too weak a requirement of autonomy for Lindley, in that "it gives insufficient weight to the role of the agent as a deliberator" who can revise and question the makeup of her desires and ends (p. 43). However, Lindley does think that rationality is a proper condition of autonomy. The kind of rationality that he has in mind, which he derives from Mill and labels "active theoretical rationality," requires that an agent take an

active role in the investigation of the truth of her beliefs and the validity of her desires (pp. 63–70). What all of these views have in common, which they have inherited in different degrees from Kant, is the demand that agents display some capacity for reasoning in order to be autonomous.

However, demanding that the conditions for autonomy include a rationality requirement of any sort, such that a person is only autonomous if she displays some level of cognitive expertise in the analysis of her desires and values, raises two crucial questions. Does the inclusion of a rationality requirement of these sorts render the characteristic of autonomy unacceptably indeterminate? And, does requiring rationality for autonomy essentially conflict with the core idea of *self*-government that autonomy is meant to express? Doubts of the first sort can be expressed this way: any model of autonomy that demands that the autonomous agent be rational will, by that token, imply that agents with varying degrees of decision-making competence will then possess the property of autonomy to differing degrees. While this does not seem problematic in itself, it seems to conflict with the intuition that autonomy is a property that can be valued equally in one's treatment of individual persons. For instance, if respect for autonomy is the basis for a general antipaternalism, then the "sliding scale" conception of autonomy would, strictly speaking, allow differing degrees of paternalistic intervention according to the level of competence a person displays in decision making. Should paternalism be less inconsistent with autonomy when the person interfered with is not as smart as another? (For a similar point, cf. Brock 1983.) Moreover, when is this magic moment when a person develops the proper level of rational capabilities necessary for autonomy, and why aren't the capacities exhibited at this *threshold* noted as the true conditions for autonomy in themselves? On such a view, it seems, autonomy is a vague, developmental property that no one truly instantiates and a notion too weak to support the further normative requirement that autonomy must be respected equally in all persons.

The second worry about rationality and autonomy is related to Berlin's attack on the notion of positive liberty (where, as he assumed, that notion included normative conditions of rationality and value). He points out that "once I take this view [that freedom requires that I act in accordance with the demands of reason], I am in a position to ignore the actual wishes of men or societies, to bully, oppress, torture them in the name . . . of their 'real' selves, in the secure knowledge that whatever is the true goal of man (happiness, the performance of duty . . .) must be identical with his freedom—the free choice of his 'true,' albeit often submerged and inarticulate, self" (Berlin 1969, p. 133).

Young (1986) harbors similar reservations about demanding that an agent be rational to be autonomous. Although he claims that "not being seriously irrational does seem to be necessary for autonomy," he urges that "there are ways of exercising autonomy . . . which do not depend on a highly developed capacity for calculation, penetrating logical analysis or any of the other trappings of the forensic approach to reasoning" (p. 12). Hence an answer to the question of whether (and to what extent) an agent must be rational to be autonomous has not been settled on in the literature.

II. AUTONOMY IN MORAL THEORY

Autonomy and Utility

I now wish to survey the ways that autonomy is dealt with in the context of moral theory. First, Utilitarianism: Utilitarianism does not usually spring to mind within

a discussion of the nature and value of autonomy. Despite the fact that Mill placed such a high value on individuality ("as one of the elements of well-being"), Utilitarians have often been criticized for a lack of emphasis on such things as autonomy, individuality, "the separateness of persons," and the like. In recent treatments of this topic that focus specifically on autonomy, Haworth (1984 and 1986) and Elster (1982 and 1983) have re-raised the charge that Utilitarianism is seriously remiss in its inability to give a direct account of the value of autonomy. While Elster argues that Utilitarianism is incomplete in this regard, Haworth claims that indeed if Utilitarianism is faithful to its own motivating assumptions, it is in fact committed to attributing a nonderivative value to autonomy.

The locus of this question need not be Utilitarianism as a whole, but rather the component of the theory known as welfarism. Although there are numerous variations in accounts of welfare or utility, welfa*rism* is a claim about human value which says that any state of affairs X can be exhaustively evaluated with no more information than that concerning the welfare (or utility) levels of the individuals in X (cf. Sen 1979, p. 471). All welfare has value and only welfare has value; all other things thought valuable must be reducible to welfare. Hence, it is not just Utilitarianism which stands or falls along with the plausibility of welfarism as a theory of value, but any view that rests on such plausibility, such as the distributive principle of "equality of welfare." Of course the concept of human welfare admits of a variety of interpretations, ranging from an extreme subjective, noncomparable measure, to more idealized interpretations. Also, the familiar dichotomy between so-called mental-state conceptions of welfare and desire-satisfaction views must be kept in mind. For all of these, though, the question looms whether the autonomy of a person, or of a person's preferences, fits into the nexus of value in relation to which the theory directs moral agents to act.

Welfarists can try to take account of autonomy in a great number of ways, either directly or indirectly. Seen as a property of whole persons, the quality of being autonomous can be accounted for in a purely instrumental fashion. A welfarist could argue that being left free to form one's own views and character, and being given the means to develop one's critical capacity for this, is the surest method to ensure happiness for oneself (cf. Haworth 1986, chap. 7; and Dworkin 1982, pp. 55–56). It has been argued, however, that Utilitarians cannot in this way take account of the *intrinsic* value of individual autonomy (or individuality) in a way that gives proper weight to the separate dignity of persons. (Cf. Rawls 1971, pp. 22 ff.)

Haworth (1986, chap. 7) argues that Utilitarianism, via the motivations that support the value foundations of the theory, has a "commitment" to the value of autonomy. The central commitment is that, at a basic intuitive level, welfare can only be considered the exclusive value in our moral universe if the preferences that constitute welfare are produced in a way consistent with the agent's autonomy. If we can imagine a person who is induced to desire some good heteronomously (via subliminal advertising, say), it is implausible to regard the person as better off when that desire becomes satisfied. Simply put, the foundational role that welfarism must play in the structure of Utilitarianism holds up only if the welfare that is to be maximized represents somehow the satisfaction of autonomous desires of the agent. As Haworth writes, "in looking at the matter in this way [i.e., welfarism] we are supposing that the individual has determined what is to count as good for him, that the preference he holds really is his, rather than one that has merely shown up. A subliminally induced desire for Coke and a brainwashed

cult member's desire to devote eighteen hours a day to selling poppies are not among the preferences one is enjoined to respect by the principles that we are to get our view of what is good-as-such by asking what people want-as-such" (1986, p. 180).

Elster argues along somewhat similar lines in spelling out various examples of "adaptive preference formation" (1982, pp. 219 ff.). The most notorious example of this is the phenomenon of "sour grapes" mentioned above, where a person's (or a fox's) inability to get what she desires effectively extinguishes her desire for the thing in question. In this way, a person's position in the distribution of goods may directly affect the preferences the person has over the goods available. This raises the general challenge to welfarism: "Why should individual want satisfaction be the criterion of justice and social choice when the individual wants themselves may be shaped by a process which preempts the choice?" (1983, p. 109). For both Haworth and Elster the argument is that welfarism is in need of a supplementary principle of "autonomous preference formation," or of "autonomous persons," to shore up the conception of value upon which the theories in question rest.

While these considerations bear most heavily on purely subjective conceptions of welfare, they may merely point to the need for certain idealizing conditions for desires of the sort that Utilitarians have already seen the need to adopt. Brandt (1979), for example, in developing a mental state conception of welfare, includes conditions of rationality of desires such that desires based on false beliefs, ones produced artificially, those based on generalizations from untypical examples, and ones resulting from early deprivation need not be counted among the components of the good to be maximized in a Utilitarian moral theory (pp. 209 ff.). Whether the conditions of "cognitive psychotherapy" that Brandt suggests to cleanse the desire set of the agent to be considered in moral decisions coincide with conditions of autonomous preferences is not clear. But the move away from pure subjectivism about desires (where all preferences are viewed as having the same status) to the slightly idealized conceptions of welfare like Brandt's clears the way for a more thoroughgoing inclusion of autonomy in the specification of welfare that appears in various moral and political principles.

A slightly different attack on Utilitarianism in this regard is given by Rawls (1982). The crux of Rawls's argument is this: traditionally, the utility of an individual is seen as a function which maps preferences onto available goods. Imagine, though, that not only are actual goods put into the utility function, but also those aspects of a person that enable her to translate available goods into satisfaction. On this model we can have preferences over goods as well as over desires, natural skills, traits of character, and the like, so that whatever it is that explains differences in tastes among individuals can itself be an argument in the function that measures utility. With this in mind, we could rank the various combinations of tastes and available goods in terms of comparable utility. Hence if a person has a set of preferences—even one which is intimately connected to that person's history, upbringing, and character—she would, on a welfarist account, be made *better off* were she to trade in her desires and character for any other that will yield a higher ranking function of subjective well-being. If one resists such a recommendation it must be that these comparisons, limited as they must be to considerations of subjective welfare, leave totally out of account the special relation that obtains between a person and that person's set of (autonomous) preferences. Hence, autonomy, insofar as it represents such a relation, is not properly accounted for by welfarism.

One could respond to Rawls's argument in at least two ways. First, it is left open here whether or not this "special relation" actually amounts to autonomy—whether, for instance, it is something like the condition of "identification" that is seen as indicative of autonomy. One could certainly have a special relation to heteronomous desires, in that one admits they are not autonomous but would not trade them in for a set higher on the utility scale. A second response is that the argument has overlooked a plausible connection between autonomy and utility, one where the desire for autonomy is just that, a desire, and hence should be counted as among the individual's set of preferences along with all the others. This view sees autonomy as the *object* of a preference, and thus a component of welfare whenever the person in question has such a desire. To the extent, then, that the relation between a person and her lower order preferences is an object of special value *for that person,* then the autonomy which is expressed in that relation can be included in the elements of her well-being, and welfarism is none the worse for it. The problem with this reply is that this preference (for the autonomy of one's preferences) may not be produced autonomously itself, in which case the welfarist would have no account of any difference (vis-à-vis the person's well-being) between such a person's desires and more autonomously formed preferences.

Autonomy, Justice, and Rights

The relation between the autonomy of individuals and principles of rights and justice is a much more familiar approach to the role that autonomy plays in moral theory. Many, in fact, see the relation between autonomy and the possibility for human agency as the foundation of morality in general, and of human rights in particular. However, accounts which explicate the relation between autonomy and principles of justice and rights should be separated into two differing categories. On the one hand, writers like Rawls (1971, 1980) and Gewirth (1978)—both heavily influenced by Kant—see autonomy as one principle property of persons that determines their ability to derive the principles of morality and justice. This amounts to construing autonomy as a kind of moral neutrality by which agents can be said to construct or derive moral principles from *their* point of view. (Cf. Hill 1987, pp. 131–33, for a discussion of this construal of autonomy.) Hence, for Rawls, individuals in the original position display "rational autonomy" and are impartial in not being guided by prior conceptions of justice; and being ignorant of any personal characteristics (race, religion, talents, place in society, and their own conception of the good) they are not biased in their judgment concerning the correct principles of justice. In this way, agents achieve "full autonomy" when they go on to act (in a well-ordered society out from behind the veil of ignorance) in accordance with principles that are "self-imposed" in the sense that they would have been chosen (by the agents) under fair and neutral conditions.

This can be contrasted with the view that autonomy is a valuable character trait of individuals which, from the point of view of the moral *theorist,* should be afforded respect and/or protection. Although they come to very different conclusions about rights, I take Richards (1981) and Nozick (1974, chap. 3) to exemplify this method. Autonomy, on Richards's view, when it is distinguished properly from the related but misleading ideas associated with it, provides the basis for the right to be treated as a free and equal moral person, a fundamental human right. (Scanlon 1972 uses a similar method in deriving the right to freedom of expression from the basic value of autonomy.) Nozick's well-discussed defense

of libertarian rights (absolute side constraints on the actions of others) also rests on the basic value of what he calls "the separateness of persons." This amounts to something very close to autonomy since, Nozick argues, the characteristics of persons that warrant the root idea that they are owed basic respect (no one can be used as a resource for another) is their capacity for rational agency and the ability to formulate a plan of life (1974, pp. 48–51). For both Richards and Nozick, autonomy is seen as the characteristic of person's whose value (from the point of view of a theorist) is only protected properly when a certain set of human or natural rights is attributed to the agent and respected by others.

The Value of Autonomy

Building a moral theory on the foundational value of autonomy raises the issue of what exactly that value issues in. There are many approaches to the question of the value of individual autonomy, corresponding, at least, to the various construals of the concept itself. The above discussion of the relation between autonomy and welfare in many ways covers much of this ground, since human welfare simply is, for many, *the* central value. One difficulty, however, in attributing autonomy noninstrumental and objective value in all cases of its presence is that our judgments about its value in some cases may depend on the kind of choices being made by the autonomous person. Do we want to say that the autonomy of thoroughly evil acts adds to the value of those acts in any way? It might seem just the opposite, that the fact that the nonautonomous evil act is by that token better than the same act performed autonomously (cf. Young 1982, p. 43).

Despite these worries, many would want to claim that autonomy is a fundamental human value, whose presence is constitutive of human agency itself (cf. Hurka 1987). Perhaps the most clear and straightforward discussion of the value of autonomy occurs in Young (1982). He distinguishes between the possibility that autonomy has intrinsic value from the claim that it is of only instrumental value (insofar as it is conducive, in particular, to happiness). The latter view is famously attributed to Mill, but Young argues that Mill came much closer to the "intrinsic value" claim. Young also argues that the instrumental view is implausible on its face since one can imagine *Brave New World*–type cases in which people are made 'happy' despite their lack of autonomy, in which case autonomy is not necessary for happiness and thus cannot be valuable on the basis of such necessity. Young goes on to argue that the intrinsic value of autonomy can be found in its relation to human agency itself and individual self esteem. "Autonomy . . . is the means to our working out our projects in the world. In exercising it, in being self-directing we make our lives . . . our *own,* and this is conducive to self esteem" (Young 1982, p. 43).

Now Young's explication of the value of autonomy must be carefully described. For if by "conducive to self-esteem," he means *causally* conducive, then the value of autonomy will turn out to be instrumental after all. What must be said here is that autonomy is somehow *constitutive* of self-esteem, or moral agency, or the like, and as such is part of the value of these things. Indeed a Utilitarian, of a complex sort, could argue that happiness simply means the satisfaction of "autonomously formed" desires (hence ruling out "experience machine" and *Brave New World* kinds of counterexamples); and the autonomy of the desires constitutes the value of their satisfaction for the agent. Autonomy, on this view, does not *produce* other things of value, it is *part* of those things.

One could perhaps argue in a more general way that the postulation that anything *whatever* is of value to persons—freedom, welfare, dignity—presupposes that autonomy of agents is also of value. For all but the most purely objective theories of value, part of the claim that something X has value for people is that people (under some conditions) in fact value it. This latter type of claim could only be plausible if it is presupposed that such people value X *autonomously*—that the judgment of its value is a reflection of the agent's authentic self. Any claim that some X has value for people must presuppose that the evaluations of a person's own interests and needs, according to which that X has value, are themselves autonomous. Therefore autonomy will be constitutive of any specification of value. No claim of value could be plausible unless it is presupposed that such judgments about what is good for people are autonomously made.

Conclusion

I have chosen for the focus of this survey the role that autonomy plays in abstract questions of moral theory, but of course much ground-breaking work on the concept is often done in the context of other issues. The question of the justification of paternalism, for example, comes immediately to mind. (Cf. Sartorius 1983 for a recent installment in this literature.) In the realm of medical ethics, it is often asked if certain treatment (or lack of treatment) will or will not infringe on the autonomy of the patient. (Cf. e.g., Miller 1981 and Edwards 1981.) Also, questions about autonomy are central to issues in feminism, where the effects of living in a patriarchal society are examined in relation to the possibility of autonomy for women (cf. Meyers 1987*a*, 1987*b*); add to this the area of educational theory which surrounds the question of which pedagogical strategies effectively promote the development of autonomy in the student (cf. Sanders 1981; Dearden 1972; and Downie and Telfer 1971).

So this survey, despite its length, is far from exhaustive,[7] even at the rather abstract theoretical level that was chosen as its focus. The major purpose here was to survey those interrelated discussions in the recent political and philosophical literature that attempted to work out a coherent conception of individual autonomy, and to explain the relation autonomy has with other values and principles in moral theories. What this at least shows is that the notions of self-government that autonomy is meant to express do amount to a citadel, but one whose construction is still in progress. What no one denies is that further refinement of the structure is not only necessary but crucial in the philosophical analysis and defense of the central values of a free society.

REFERENCES

Arneson, Richard. 1985. Freedom and Desire. *Canadian Journal of Philosophy* 15:425–48.

Benson, John. 1983. Who Is the Autonomous Man? *Philosophy* 58:5–17.

Berlin, Isaiah. 1969. Two Concepts of Liberty. In *Four Essays on Liberty*, pp. 118–72. Oxford: Oxford University Press.

Brandt, Richard. 1979. *A Theory of the Good and the Right.* Oxford: Oxford University Press.

Braybrooke, David. 1974. From Economics to Aesthetics: The Rectification of Preferences. *Nous* 8:13–24.

Brock, Dan. 1983. Paternalism and Promoting the Good. In *Paternalism*, ed. Rolf Sartorius, pp. 237–60. Minneapolis: University of Minnesota Press.

7. For example, I have not discussed the political implications of attributing equal value to the autonomy of citizens; cf., e.g., Young (1986), chap. 8 for a discussion of this.

Christman, John. 1987. Autonomy: A Defense of the Split-Level Self. *Southern Journal of Philosophy* 25:281–93.

Dearden, R. F. 1972. Autonomy and Education. In *Education and the Development of Reason*, ed. R. F. Dearden et al., pp. 451–52. London: Routledge & Kegan Paul.

Downie, R. S., and Telfer, E. 1971. Autonomy. *Philosophy* 46:296–301.

Dworkin, Gerald. 1976. Autonomy and Behavior Control. *Hastings Center Report* 6:23–28.

Dworkin, Gerald. 1981. The Concept of Autonomy. In *Science and Ethics*, ed. R. Haller, pp. 203–13. Amsterdam: Rodopi, 1981.

Dworkin, Gerald. 1982. Is More Choice Better Than Less? *Midwest Studies in Philosophy*, vol. 7, ed. Peter French et al., pp. 47–62. Minneapolis: University of Minnesota Press.

Edwards, Rem. 1981. Mental Health as Rational Autonomy. *Journal of Medicine and Philosophy* 6:309–21.

Elster, Jon. Sour Grapes—Utilitarianism and the Genesis of Wants. In Sen and Williams, eds. 1982, pp. 219–38.

Elster, Jon. 1983. *Sour Grapes*. Cambridge: Cambridge University Press.

Feinberg, Joel. 1973. *Social Philosophy*. Englewood Cliffs, N.J.: Prentice-Hall.

Feinberg, Joel. 1986. *Harm to Self*. Vol. 3 of *The Moral Limits of the Criminal Law*. New York: Oxford University Press.

Frankfurt, Harry. 1971. Freedom of the Will and the Concept of the Person. *Journal of Philosophy* 68:829–39.

Frankfurt, Harry. 1987. Identification and Wholeheartedness. In *Responsibility, Character and the Emotions*, ed. F. Schoeman, pp. 27–45. Cambridge: Cambridge University Press.

Friedman, Marilyn. 1986. Autonomy and the Split-Level Self. *Southern Journal of Philosophy* 24:19–35.

Gewirth, Alan. 1978. *Reason and Morality*. Chicago: University of Chicago Press.

Haworth, Lawrence. 1984. Autonomy and Utility. *Ethics* 95:5–19.

Haworth, Lawrence. 1986. *Autonomy: An Essay in Philosophical Psychology and Ethics*. New Haven, Conn.: Yale University Press.

Hill, Thomas. 1984. Autonomy and Benevolent Lies. *Journal of Value Inquiry* 18:251–67.

Hill, Thomas. 1987. The Importance of Autonomy. In *Women and Moral Theory*, ed. E. Feder Kitan and D. Meyers, pp. 129–38. Totowa, N.J.: Rowman & Littlefield.

Hurka, Thomas. 1987. Why Value Autonomy? *Social Theory and Practice* 13:361–82.

Kant, I. 1969. *Foundations of the Metaphysics of Morals*, with critical essays. Trans. Lewis White Beck, ed. R. P. Wolff. New York: Bobbs-Merrill.

Lindley, Richard. 1986. *Autonomy*. London: Macmillan.

MacCallum, Gerald. 1967. Negative and Positive Freedom. *Philosophical Review* 76:312–32.

Meyer, Michael J. 1987. Stoics, Rights and Autonomy. *American Philosophical Quarterly* 24: 267–71.

Meyers, Diana. 1987*a*. Personal Autonomy and the Paradox of Feminine Socialization. *Journal of Philosophy* 84:619–28.

Meyers, Diana. 1987*b*. The Socialized Individual and Individual Autonomy: An Intersection between Philosophy and Psychology. In *Women and Moral Theory*, ed. E. Feder Kitan and D. Meyers, pp. 139–53. Totowa, N.J.: Rowman & Littlefield.

Miller, Bruce. 1981. Autonomy and the Refusal of Life-saving Treatment. *Hastings Center Report* 11:22–28.

Neely, Wright. 1974. Freedom and Desire. *Philosophical Review* 83:32–54.

Nozick, Robert. 1974. *Anarchy, State and Utopia*. New York: Basic Books.

Rawls, John. 1971. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Rawls, John. 1980. Kantian Constructivism in Moral Theory. *Journal of Philosophy* 77: 515–35.

Rawls, John. 1982. Social Unity and Primary Goods. In Sen and Williams, eds., 1982.

Richards, David. 1981. Rights and Autonomy. *Ethics* 92:3–20.

Sartorius, Rolf. 1983. *Paternalism*. Minneapolis: University of Minnesota Press.

Sanders, Steven. 1982. Autonomy, Authority and Moral Education. *Journal of Social Philosophy* 13:18–24.
Scanlon, T. M. 1972. A Theory of Freedom of Expression. *Philosophy and Public Affairs* 1: 204–26.
Sen, Amartya. 1977. Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy and Public Affairs* 6:317–44.
Sen, Amartya. 1979. Utilitarianism and Welfarism. *Journal of Philosophy* 76:463–89.
Sen, Amartya, and Williams, Bernard, eds. 1982. *Utilitarianism and Beyond.* Cambridge: Cambridge University Press.
Strawson, Galen. 1986. *Freedom and Belief.* Oxford: Oxford University Press.
Thalberg, Irving. 1978. Hierarchical Analyses of Unfree Action. *Canadian Journal of Philosophy* 8:211–26.
Thalberg, Irving. 1979. Socialization and Autonomous Behavior. *Tulane Studies in Philosophy* 28:21–37.
Watson, Gary. 1975. Free Agency. *Journal of Philosophy* 72:205–20.
Watson, Gary. 1987. Free Action and Free Will. *Mind* 96:145–72.
Yaari, Menahem. 1977. Endogenous Changes in Tastes: A Philosophical Discussion. *Erkenntnis* 11:157–96.
Young, Robert. 1980. Autonomy and the Inner Self. *American Philosophical Quarterly* 17: 35–43.
Young, Robert. 1982. The Value of Autonomy. *Philosophical Quarterly* 32:35–44.
Young, Robert. 1986. *Personal Autonomy: Beyond Negative and Positive Liberty.* New York: St. Martin's Press.

ADDITIONAL READINGS

Allen, R. T. 1982. Rational Autonomy: The Destruction of Freedom. *Journal of Philosophy of Education* 16:199–207.
Beardsley, Elizabeth. 1971. Privacy: Autonomy and Selective Disclosure. In *NOMOS,* vol. 8, *Privacy,* ed. R. Pennock and J. Chapman, pp. 56–70. New York: Atherton Press.
Benn, S. I. 1967. Freedom and Persuasion. *Australasian Journal of Philosophy* 45:259–75.
Benn, S. I. 1976. Freedom, Autonomy and the Concept of a Person. *Proceedings of the Aristotelian Society* 66:109–30.
Benn, S. I. 1982. Individuality, Community and Autonomy. In *Community,* ed. E. Kamenka. London: E. Arnold Publishers.
Benson, Paul. 1987. Freedom and Value. *Journal of Philosophy* 84:465–87.
Bernstein, M. 1983. Socialization and Autonomy. *Mind* 92:120–23.
Crocker, Lawrence. 1980. *Positive Liberty.* The Hague: Martinus Nijhoff.
Dearden, R. F. 1975. Autonomy as an Educational Ideal I. In *Philosophers Discuss Education,* ed. S. C. Brown. London: Macmillan.
Dworkin, Gerald. 1970. Acting Freely. *Nous* 4:367–83.
Dworkin, Gerald. 1978. Moral Autonomy. In *Morals, Science and Sociality,* ed. H. Tristram Englehard, Jr., and Daniel Callahan. New York: Hastings Center.
Engstrom, Stephen. 1988. Conditioned Autonomy. *Philosophy and Phenomenological Research* 48:435–53.
Fleming, N. 1981. Autonomy of the Will. *Mind* 90:201–23.
Goodin, Robert. 1981. The Political Theories of Choice and Dignity. *American Philosophical Quarterly* 18:91–100.
Hill, Sharon. 1975. Self Determination and Autonomy. In *Today's Moral Problems,* ed. Richard Wasserstrom. New York: Macmillan.
Kuflick, Arthur. 1984. The Inalienability of Autonomy. *Philosophy and Public Affairs* 13: 271–98.
Ladenson, R. F. 1975. A Theory of Personal Autonomy. *Ethics* 86:30–48.

Nauta, Lolle. 1985. Historical Roots of the Concept of Autonomy in Western Philosophy. *Praxis* 4:363–77.

Parent, W. D. 1974. Some Recent Work on the Concept of Liberty. *American Philosophical Quarterly* 11:149–67.

Rachels, James. 1986. *The Ends of Life: Euthanasia and Morality.* Oxford: Oxford University Press.

Richards, David. 1979. Sexual Autonomy and the Constitutional Right to Privacy: A Case Study in Human Rights and the Unwritten Constitution. *Hastings Law Journal* 30: 957–1018.

Rorty, Amelie, ed. 1976. *The Identities of Persons.* Berkeley: University of California Press.

Schrader, George A. 1963. Autonomy, Heteronomy, and Moral Imperatives. *Journal of Philosophy* 60:65–77.

Siegler, Mark. 1977. Critical Illness: The Limits of Autonomy. *Hastings Center Report* 8: 12–15.

Taylor, Charles. 1976. Responsibility for Self. In Rorty, ed., 1976, pp. 281–99.

Wolf, Susan. 1987. Sanity and the Metaphysics of Responsibility. In *Responsibility, Character and the Emotions,* ed. F. Schoeman. Cambridge: Cambridge University Press.

Wolff, Robert Paul. 1979. *In Defense of Anarchism.* New York: Harper & Row.

Young, Robert. 1979. Compatibilism and Conditioning. *Nous* 13:361–78.

Young, Robert. 1980. Autonomy and Socialization. *Mind* 89:565–76.