National Library of Canada
Acquisitions and Bibliographic Services Branch
395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale du Canada
Direction des acquisitions et des services bibliographiques
395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

PERSONAL AUTONOMY

by

Susan Dimock

Submitted in partial fulfilment of the requirements
for the degree of Doctorate of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
February 1994

ISBN 0-315-93770-X

Canada

Name Susan Ann Dimock

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

Personal Autonomy - Philosophy (SUBJECT TERM) 0422 (SUBJECT CODE) U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS
Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION
General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523

Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Special 0529
Special Subjects 0530
Student Counseling 0710
Teacher Training 0288
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS
Language
General 0679
Ancient 0289
Linguistics 0290
Modern 0291
Literature
General 0401
Classical 0294
Comparative 0295
Medieval 0297
Modern 0298
African 0316
American 0591
Asian 0305
Canadian (English) 0352
Canadian (French) 0355
English 0593
Germanic 0311
Latin American 0312
Middle Eastern 0315
Romance 0313
Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY
Philosophy 0422
Religion
General 0318
Biblical Studies 0321
Clergy 0319
History of 0320
Philosophy of 0322
Theology 0469

SOCIAL SCIENCES
American Studies 0323
Anthropology
Archaeology 0324
Cultural 0326
Physical 0327
Business Administration
General 0310
Accounting 0272
Banking 0770
Management 0454
Marketing 0338
Canadian Studies 0385
Economics
General 0501
Agricultural 0503
Commerce-Business 0505
Finance 0508
History 0509
Labor 0510
Theory 0511
Folklore 0358
Geography 0366
Gerontology 0351
History
General 0578

Ancient 0579
Medieval 0581
Modern 0582
Black 0328
African 0331
Asia, Australia and Oceania 0332
Canadian 0334
European 0335
Latin American 0336
Middle Eastern 0333
United States 0337
History of Science 0585
Law 0398
Political Science
General 0615
International Law and Relations 0616
Public Administration 0617
Recreation 0814
Social Work 0452
Sociology
General 0626
Criminology and Penology 0627
Demography 0938
Ethnic and Racial Studies 0631
Individual and Family Studies 0628
Industrial and Labor Relations 0629
Public and Social Welfare 0630
Social Structure and Development 0700
Theory and Methods 0344
Transportation 0709
Urban and Regional Planning 0999
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES
Agriculture
General 0473
Agronomy 0285
Animal Culture and Nutrition 0475
Animal Pathology 0476
Food Science and Technology 0359
Forestry and Wildlife 0478
Plant Culture 0479
Plant Pathology 0480
Plant Physiology 0817
Range Management 0777
Wood Technology 0746
Biology
General 0306
Anatomy 0287
Biostatistics 0308
Botany 0309
Cell 0379
Ecology 0329
Entomology 0353
Genetics 0369
Limnology 0793
Microbiology 0410
Molecular 0307
Neuroscience 0317
Oceanography 0416
Physiology 0433
Radiation 0821
Veterinary Science 0778
Zoology 0472
Biophysics
General 0786
Medical 0760

EARTH SCIENCES
Biogeochemistry 0425
Geochemistry 0996

Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleoecology 0426
Paleontology 0418
Paleozoology 0985
Palynology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES
Environmental Sciences 0768
Health Sciences
General 0566
Audiology 0300
Chemotherapy 0992
Dentistry 0567
Education 0350
Hospital Management 0769
Human Development 0758
Immunology 0982
Medicine and Surgery 0564
Mental Health 0347
Nursing 0569
Nutrition 0570
Obstetrics and Gynecology 0380
Occupational Health and Therapy 0354
Ophthalmology 0381
Pathology 0571
Pharmacology 0419
Pharmacy 0572
Physical Therapy 0382
Public Health 0573
Radiology 0574
Recreation 0575

Speech Pathology 0460
Toxicology 0383
Home Economics 0386

PHYSICAL SCIENCES
Pure Sciences
Chemistry
General 0485
Agricultural 0749
Analytical 0486
Biochemistry 0487
Inorganic 0488
Nuclear 0738
Organic 0490
Pharmaceutical 0491
Physical 0494
Polymer 0495
Radiation 0754
Mathematics 0405
Physics
General 0605
Acoustics 0986
Astronomy and Astrophysics 0606
Atmospheric Science 0608
Atomic 0748
Electronics and Electricity 0607
Elementary Particles and High Energy 0798
Fluid and Plasma 0759
Molecular 0609
Nuclear 0610
Optics 0752
Radiation 0756
Solid State 0611
Statistics 0463

Applied Sciences
Applied Mechanics 0346
Computer Science 0984

Engineering
General 0537
Aerospace 0538
Agricultural 0539
Automotive 0540
Biomedical 0541
Chemical 0542
Civil 0543
Electronics and Electrical 0544
Heat and Thermodynamics 0348
Hydraulic 0545
Industrial 0546
Marine 0547
Materials Science 0794
Mechanical 0548
Metallurgy 0743
Mining 0551
Nuclear 0552
Packaging 0549
Petroleum 0765
Sanitary and Municipal 0554
System Science 0790
Geotechnology 0428
Operations Research 0796
Plastics Technology 0795
Textile Technology 0994

PSYCHOLOGY
General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

TABLE OF CONTENTS

ABSTRACT

This dissertation provides a philosophical analysis of personal autonomy. Personal autonomy is defined as the condition of being self-directed. The conditions which make such self-direction possible are then explored.

Self-direction requires that one's actions are motivated by authentic reasons for action. One makes some of one's desires authentic by critically reflecting upon and identifying with them. One identifies with a particular desire when one approves of it as a reason for action, thus desiring that it be an effective desire. Provided that one's identification with a desire is decisive and one is not ambivalent with respect to it, identification is sufficient for the authenticity of one's desires.

This capacity to adopt authentic reasons for action cannot be sufficient for the autonomy of one's desires or actions, though. For authenticity is a function solely of the psychological states of an individual; yet autonomy cannot be adequately explicated solely by reference to an agent's subjective states. I refer to those theories which make the autonomy of a person's desires solely a function of her psychological states as "internalist", and argue that such theories must be rejected. An "externalist" is one who denies that autonomy is wholly a function of the psychological states of the individual, and so holds that the autonomy of a desire is determined, at least in part, by facts which are independent of (external to) the subjective attitudes of the agent.

I defend a form of externalism, which makes the autonomy of a desire depend upon the following conditions:

1) the agent must approve of it as a reason for action; 2) the agent must have the capacity to respond appropriately to whatever (objective) reasons there are for and against it; 3) the agent must have been able to avoid falling into error (both evaluative and nonevaluative) concerning the object of the desire and the desire itself; 4) the agent's approval of it is not caused solely by a restriction of her feasible options or coercion.

I argue that this is a more plausible theory than other externalist positions, and it can meet the objections which have beset internalist theories.

CHAPTER I

PERSONAL AUTONOMY AND AUTHENTICITY

I.1 INTRODUCTION

Personal autonomy consists, in its most basic sense, in the condition of being self-directed or self-governed. This understanding of the concept is neither startling nor innovative; it is implied by the etymology of the word and is shared by most philosophers who discuss it. Personal autonomy, understood as self-direction, is to be contrasted both with direction by another/others and direction by unreflective passion, impulse or desire.[1]

In virtue of what properties are persons able to direct themselves in what they do? This is a subject about which philosophers have disagreed at length. What seems clear, however, is that any adequate explication of the concept of personal autonomy may be expected to capture the insistence, found within diverse theories of that condition, that autonomy consists of action which reflects one's "true" or "authentic" self. This core idea can be gleaned in the following familiar expressions: the autonomous man is "his own person"; autonomous agents are "authentic"; an autonomous agent acts from desires and values which are properly seen as "her own"; the actions of autonomous agents reflect "who they really are";

---

[1] These are not the only forms which heteronomy or lack of autonomy can take, of course, but manipulation and wantonness, both broadly construed, capture many important cases of lack of personal autonomy in motivation.

an autonomous agent acts "according to her own lights", "marches to his own beat", and the like. What all of these have in common is the idea that there is an authentic self, comprised of the beliefs, desires and values which are truly the agent's own, and that autonomous behaviour flows from and expresses that true self. I shall summarize these intuitions in saying that autonomous persons are authentic in their reasons for action.

The general phrase "reason for action" is ambiguous in an important sense, however; for there is a familiar usage in which one's reasons for action can be non-authentic. Thus we might speak of a hypnotically induced desire as the reason why Jill stole the blueprints, or of the addictive desire for heroin as the reason why Bob took the drug. Such reasons for action as these are purely explanatory reasons; they may be sufficient in explanations as "reasons why" someone performed some action. Any desire may provide such a reason for action, insofar as the desire, plus the relevant belief that some action will bring about the desired state of affairs, can explain the action. But reasons for action in this sense would also include such things as (the effects of) brain tumours, hormonal imbalances, subliminal conditioning, a demonic neuro-scientist who manipulates one's brain, an epileptic fit and the like; all of these things could explain the behaviour of an individual on a given occasion. Likewise, the desire for food can function as an explanation of why the infant cries, or the desire for affection as why the dog learns the tricks its master teaches it. We can offer purely explanatory reasons for the actions of even non-autonomous entities.

Theories of evolutionary advantage are ripe with explanations of why creatures act the way they do, though there is no assumption that either the creatures or their actions are autonomous. Such explanations do not cite the agent's authentic reasons for action. Authentic reasons for action are "the person's reasons" for acting. They are both explanatorily adequate and defensible by the agent whose reasons they are. They are reasons which the agent would cite both to explain and justify her action.[2] Autonomous agents act on reasons which are authentic in this sense.

This account of authenticity immediately raises a pressing question: how does one acquire authentic reasons for action? Given that we acquire many beliefs, desire, values and projects uncritically from others, this is an important question for anyone who wishes to argue that autonomy is attained through the development of authentic reasons for action. Insofar as our reasons for action are acquired and maintained uncritically from others -- whether particular others or our social environment more generally is irrelevant -- they cannot plausibly be described as authentic. We must, then, be able to make some of our reasons for action our own; but how?

One answer to this question has received considerable attention in recent years. It is, roughly, that one makes one's pre-reflective desires, beliefs, projects and values authentic by critically reflecting upon them and identifying with some

---

[2] Stephen Darwall has drawn a similar distinction concerning the various meanings which the phrase "reasons for action" might have to that which I employ here. Cf. Impartial Reason (Cornell University Press, 1983), p. 28.

while rejecting others. Those which are thus identified with become the agent's authentic reasons for acting, and actions which issue from such reasons are autonomous. This, I believe, is the core idea to be found in the "bi-level" theories of autonomy developed by Gerald Dworkin, Harry Frankfurt and Lawrence Haworth.

I.2 BI-LEVEL THEORIES OF AUTONOMY

On the bi-level view of autonomy the basic capacity which makes autonomy possible is that of being able to reflect upon one's desires, beliefs, plans and values as possible reasons for action. In discussing this reflective activity and endorsement of one's desires, plans and values, Dworkin, Frankfurt and Haworth all employ a bi-level account of desires. The distinction which is central to bi-level accounts of autonomy is that between first-order and second-order desires. First-order desires are desires to do or not to do something; they have actions as their objects (or state of affairs which can be brought about by one's actions). Second-order desires are desires to have or not to have certain first-order desires; their objects are first-order desires. Second-order desires which are desires that some first-order desire be effective in moving one to action are called volitions.[3]

---

[3] Here I am following Frankfurt's terminology, as do most commentators in the literature on the bi-level theory of autonomy. I shall speak, pleonastically, of second-order volitions only to remind the reader of the level at which they are operative. Frankfurt calls first-order desires which are effective the agent's "will". As Richmond Campbell has pointed out, though, this is troublesome, for it creates

The relevance of the bi-level theory of desires to personal autonomy follows directly from this distinction: when we act from our higher-order desires we act on motives which we have examined and endorsed, i.e., we act on authentic desires, and thus we are self-directed. (This way of presenting the bi-level view is, actually, over-stated, for acting on a higher-order desire is only a necessary but not a sufficient condition of personal autonomy on any plausible version of the theory. Likewise, the authenticity of one's desires is only necessary for their being autonomous, but it is not sufficient.)

The intuition which drives this approach is that a person who acts from authentic motives, i.e., those which she has reflected upon and approved of as reasons for action, is self-directed in a way that is not true of someone who acts from motives which she has not examined, or which she wishes not to act on. Employment of this bi-level account of desires allows us to distinguish desires which are properly attributable to an agent as authentic from those which are not. Such an account is useful in helping to identify those motives which exemplify the agent's self-conception, and so they assist us in delineating those motives

---

problems in understanding how weakness of will (acting contrary to one's will) could be possible, though Frankfurt thinks it is possible. Campbell argues that it is better to understand the will as being at the level of second-order volitions, and he is clearly right. The remark that it is pleonastic to call volitions second-order is not meant to beg the question between Frankfurt and Campbell concerning whether to identify the agent's will at the first or second level. See Harry Frankfurt, "Freedom of the Will and the Concept of a Person" Journal of Philosophy 68 (1971); reprinted in his The Importance of What We Care About (Cambridge University Press, 1988); and Richmond Campbell, Self-Love and Self-Respect: A Philosophical Study of Egoism (Canadian Library of Philosophy, 1979), pp. 143-155.

which, when causally effective, result in acts of self-determination. For acts of self-determination are those which are motivated by authentic reasons for action.

In his important essay on "Freedom of the Will and the Concept of a Person", Frankfurt argues that what distinguishes persons from non-persons is the ability to reflect upon their desires and motives and form high-order desires and motives which have these as their objects.[4]

> Besides wanting and choosing and being moved to do this or that, men may also want to have (or not to have) certain desires and motives. They are capable of wanting to be different, in their preferences and purposes, from what they are.[5]

This reflective capacity distinguishes persons from all other kinds of beings. As reflective beings we can form desires about our own first-order desires; we can want our will (those first-order desires which are effective in moving us to action) to be different from what it is. Alternately, we can be happy that it is what it is. It is this capacity for reflection, the capacity to become critically aware of our own wills and of forming volitions of the second-order, which makes the attainment of personal autonomy possible.[6]

---

[4] It is an open question whether only human beings are capable of personhood on this view.

[5] Frankfurt, "Freedom of the Will and the Concept of a Person" Journal of Philosophy 68 (1971), p. 12; reprinted in The Importance of What We Care About. All page references to Frankfurt 1971 are from the latter collection.

[6] Harry Frankfurt, "Identification and Wholeheartedness", Responsibility, Character, and the Emotions, ed. Ferdinand Schoeman (Cambridge University Press, 1987); reprinted in The Importance of What We Care About, pp. 164-165. All citations of Frankfurt 1987 refer to the latter collection.

This capacity for critical reflection and the formation of volitions makes personal autonomy possible because in forming second-order volitions the person "identifies" himself with a first-order desire and withdraws from other first-order desires. In so doing, the desire identified with becomes "internal" to the person whose desire it is, while those which are withdrawn from become "external to" or "alien from" the person.[7] Identification makes some desires more truly "his" than others which he has at the first-order. Those desires with which one identifies are one's authentic reasons for action. And a person is self-directed when he acts on desires which are authentic in just this sense: they are internal to him and truly his own.

Frankfurt offers the example of an unwilling drug addict to explain the role of identification in his theory of freedom of the will. This addict has a desire to take the drug to which he is addicted and a desire not to take the drug, both of which are first-order desires. But the addict has also made a commitment about which of these desires he wants to be his will (his effective desire). He has decided that he wants his desire not to take the drug to be his will, i.e., he wants his desire to resist to be effective. In making this decision the addict has formed

---

[7] The contrast between desires that are "internal" to a person and those which are "external" to her is supposed to mark the difference between authentic and inauthentic desires. Internal desires are authentic, truly "hers", while external desires are hers only in some superficial sense. The use of "internal" and "external" here follows that of Dworkin and Frankfurt. As it is used here, though, internality or externality is a property of desires. I use the terms "internalist" and "externalist" later, in Chapter IV, to mark what I take to be an important division among theories of autonomy. I note this only to avoid confusion later.

a second-order volition. But he has also identified himself with one of his conflicting first-order desires and withdrawn himself from the other.[8] Before forming the second-order volition as a way of resolving the conflict among his desires, it could be said that he wanted both to take and to not take the drug. After he has identified with the desire not to take the drug, however, he can say that what he really wants to do is to resist the desire to take the drug, even though both the first-order desires may continue to be elements in his mental history and even if he succumbs to the desire for the drug at some future time.

Gerald Dworkin has written extensively in defence of a bi-level theory of autonomy which shares many important features with that offered by Frankfurt. In his original statement of the bi-level conception of autonomy, Dworkin also argued that one's higher-order endorsement of some of one's first-order desires conferred authenticity on them in virtue of the agent "identifying" with them.

> A person may identify with the influences [i.e., the first-order desires] that motivate him, assimilate them to himself, view himself as the kind of person who wishes to be moved in particular ways. Or, he may resent being motivated in certain ways, be alienated from those influences, prefer to be the kind of person who is motivated in different ways.[9]

Through the formation of second-order volitions, one identifies with some of one's first-order desires as reasons for action and withdraws from others. In

---

[8] Frankfurt 1971, p. 18.

[9] Gerald Dworkin, The Theory and Practice of Autonomy (Cambridge University Press, 1988), p. 15.

doing so one decides not only how one wants to act but the kind of person one wants to be.

> My suggestion is that it is the broader notion of autonomy that is linked with the identification of a person with his projects, values, aims, goals, desires, and so forth. It is only when a person identifies with the influences which motivate him, assimilates them to himself, views himself as the kind of person who wishes to be moved in particular ways, that these influences are to be identified as "his". If, on the contrary, a person resents being motivated in certain ways, is alienated from these influences, would prefer to be the kind of person who is motivated in different ways, then these influences, which may be causally effective, are not viewed by him as "his".[10]

Identification with one's desires, then, is the act by which an agent incorporates some desires within her self-conception and thereby makes them truly her own. If it is to play this role, however, identification must be more than mere acknowledgement. "Identification" may, in everyday usage, mean either that one merely recognizes or acknowledges a desire as something that one has, i.e., as an element in one's mental history, or it may include a component of endorsement so that by identifying with a particular desire one not only recognizes it as a component of one's self but approves of it as a part of one's self. It would be more difficult to defend that claim that the phrase "identifies with" could be understood to mean mere acknowledgement, for to say that one identifies with something seems to imply an element of endorsement (e.g., she

---

[10] Gerald Dworkin, "The Concept of Autonomy", Science and Ethics, ed. Rudolph Haller (Rodopi Press, 1981); reprinted in The Inner Citadel : Essays on Individual Autonomy, ed. John Christman (Oxford University Press, 1989), p. 60. All citations of Dworkin 1981 refer to the latter collection.

identifies with the pro-choice movement). Nonetheless, for the sake of clarity it should be noted explicitly that on bi-level theories of autonomy, identification is used in this second, normative sense.[11] Identification is not mere acknowledgement. An agent could acknowledge that she has a certain desire, say a desire to drown her bawling child in the bath,[12] yet in no way identify with it or desire that it be effective in moving her to act. She would not recognize such a desire as a component in her self-conception, and this need not be through self-deception. It is just that the desire to drown her child is no persistent part of the person she is and absolutely no part of the person she wants to be. The desire is hers, then, only in a very superficial sense. One can acknowledge that one has desires in this sense, and even recognize that those desires are characteristic of a certain type of person, without approving of those desires or seeing oneself as that sort of person. A member of the IRA, by contrast, who desires to murder a British diplomat may be expected not only to acknowledge this desire as an element in her mental history, but to approve of it, identify with it and assimilate it within her self-conception. Identification as endorsement requires that one recognize the desire in the first sense, and also that one approve of it and the kind of person which such a desire is characteristic of.

As with the ambiguity between recognizing and endorsing desires in

---

[11] This is certainly clear in Dworkin's early writings, where identification involves approval or endorsement of certain of one's first-order desires. See Dworkin 1988, pp. 15-16.

[12] I borrow this example from Gary Watson.

identification, there is a dual role which "self-conception" plays in this account. One's self-conception can mean just what sort of person one recognizes oneself as being, or the sort of person one wants to be. In the former sense one can recognize that one is miserly, say, yet one can also want to be a different sort of person, one who is more generous, perhaps. It is this second sense of self-conception, the ideal one holds up to be striven for and against which particular desires are evaluated, that is tied to the normative sense of identification. When one identifies with a particular desire one approves of it as characteristic of the kind of person one wants to be and so one wants to act on it as part of being that kind of person.

There is an objection which could arise here that should be addressed immediately. For this description of the role of identification and self-conception in the bi-level theory of authenticity might be charged with relying on a particular variant of the "wishful thinking" fallacy. For it seems to conflate the actual with the ideal. It seems to imply that to have authentic reasons for action and so be capable of acting autonomously one must think oneself (morally) perfect. Authenticity seems incompatible with the honest recognition that one has particular weaknesses and failings, and that some of one's desires are not fully defensible. This does not really follow from the view I am considering. Though we may all have some natural impulse to construe who we are as who we would like to be, it is compatible with identification's being taken in the strong normative sense I just described that one can recognize that one has some

"undesirable" desires. It is even possible that one can come to recognize that these are very ingrained features of one's personality, which frequently influence one's action. In such cases one might resign oneself to them; indeed, this might be necessary to avoid painful dissonance between one's actual and ideal self-conception or to avoid engaging in unsuccessful forms of rationalization or self-deception. Frankfurt clearly recognizes such a possibility.

> This equation of the real with the ideal does play a role in the way some people think about themselves. Nonetheless, the distinction between internal and external passions is not the same as the distinction between what is and what is not "real" in the sense of conforming to a person's ideal image of himself. Surely it is possible for a person to recognize that a certain passion is unequivocally attributable to him, even when he regrets this fact and wishes that the passion did not occur in him or move him at all. Perhaps after long struggle and disillusion with himself, a person may become resigned to being someone of whom he does not altogether approve. He no longer supposes that he is capable of bringing the course of his passions into harmony with his ideal concept of himself, and accordingly he ceases to reserve his acceptance of his passions as they are.[13]

But the identification will fail to be "wholehearted" in such cases.

Ignoring for now the difficulties posed by acts of acceptance of a desire without endorsement of it, we shall assume that identification with a desire makes the desire internal to the agent whose desire it is. On the basis of such acts of identification one develops an authentic self. This model is, then,

---

[13] Harry Frankfurt, "Identification and Externality", The Identities of Persons, ed. Amelie Rorty (University of California Press, 1977); reprinted in The Importance of What We Care About, pp. 63-64. Citations of Frankfurt 1977 refer to the latter collection.

understandably attractive, for it provides an account of the development of the authentic self -- needed for "self"-direction -- through a process of self-definition. One defines or constructs an authentic self which is constituted by those beliefs, values, projects and desires with which one identifies after critically reflecting upon them. One acts autonomously only if one's action flows from and expresses the commitments of one's authentic self.

I.3 THE INADEQUACY OF UNI-LEVEL THEORIES OF AUTONOMY

Invoking a second order of desires, as the bi-level view of authenticity does, complicates our mental ontology and so needs to be defended. For even if the bi-level theory of autonomy provides a convenient model for representing the reflective activity of self-evaluation and self-definition, such convenience may not be enough to justify adopting it unless it is clearly superior to uni-level alternatives. Many objections have been raised to the bi-level conception of authenticity precisely on the grounds that it is not necessary to invoke higher-order desires, i.e., it is claimed that we can draw the distinction between superficial desires and what a person really wants within a uni-level theory of desires. I will consider two such objections to the bi-level view, and argue that they fail to support the claim that the second-order of desire formation is not necessary for authenticity. I will offer more positive arguments in support of the claim that the ability to form second-order desires through the process of

identification is necessary for conferring authenticity upon some of our first-order desires, plans, projects and values in Chapter II.

I.3.i Are Second-Order Desires Reducible to Values?

Gary Watson has proposed a uni-level model of authenticity as an alternate to the bi-level theory which is worth exploring. In his essay, "Free Agency", Watson distinguishes between desires and values. While it is impossible to value some state of affairs without also desiring that it obtain, he claims, the converse is not true: one can desire things which one does not value or which one devalues. "The problem of free action," as Watson sees it, "arises because what one desires may not be what one values and what one most values may not be what one is finally moved to get."[14] On this view the problem of unfree action arises because of the possibility of conflict between our "motivational systems" (based on desires, appetites and passions) and our "valuation systems" (based on our judgment that a certain action is good).

> If there are sources of motivation independent of the agent's values [i.e., desires which one does not value gratifying], then it is possible that sometimes he is motivated to do things he does not deem worth doing. This possibility is the basis for the principal problem of free action: a person may be obstructed by his own

---

[14] Gary Watson, "Free Agency" Journal of Philosophy Vol. LXXII, No.8 (April 1975); reprinted in The Inner Citadel, p.112. Citations of Watson 1975 refer to the latter collection.

will.[15]

Watson posits, then, two independent sources of motivation within free agents (within agents capable of freedom of the will, that is): values and desires. Values have their origin in "the rational part of the soul" and arise from the judgment that a particular action or state of affairs is good; other desires have their origin in the appetites and passions, or in acculturalization. These latter mere desires may all exert a motivational influence which is independent of one's evaluative judgments.

This distinction between values and desires is to be marked not on the basis of the content of desires and values, and not behaviorally. Rather, the difference between values and desires is to be drawn in terms of the "attitudes" which the agent has toward them.

> We might say that an agent's values consist in those principles and ends which he -- in a cool and non-self-deceptive moment -- articulates as definitive of the good, fulfilling, and defensible life.[16]

Whatever ends one thinks good in this way are one's values, and one's values are one's authentic reasons for action.

I think there are many problems with Watson's account, not all of which are relevant here. He claims to be giving a non-hierarchical account of what one

---

[15] Watson 1975, p. 115. One might wonder here why this is the principal problem of free action, as opposed to a problem of freedom of the will. I shall ignore this difficulty here.

[16] Watson 1975, p.116

most wants to do: what one most wants to do is what one values doing. Thus the distinction between values and desires is supposed to offer a principled way of identifying those desires which are "truly the person's own" at the first-order. But this requires that only those desires which are approved of because their ends are believed to be good can be authentic; only those desires or ends which are approved of for normative reasons can be authentic, for it is the judgment that a particular principle or end is good which makes it a value. There is a great danger here of the "real" self being reduced to the "moral" self. If this account of authenticity is to form a basis for a theory of autonomy, then, it would reduce personal autonomy to moral autonomy.[17]

There are more serious worries as well. For one can have heteronomous values. One's beliefs about what is good may be the product of severe manipulation or deception. Thus one's values may be influenced in ways which undermine their authenticity. The requirement that one's judgments be made "in a cool and non-self-deceived moment" does nothing to rule out the possibility that one's values are wholly a product of socialization or acculturalization. Children appropriate the values of others, most commonly those of their parents or parent-surrogates, before they have developed the skills needed for critically assessing

[17] The contrast between personal and moral autonomy is well recognized in the literature. See Diana Meyers, Self, Society and Personal Choice (Columbia University Press, 1989), Chapter I, for details.

them.[18] There seems no reason to suppose, however, that those values are more truly their own than other desires they may have. The problem here is that values are no more necessarily internal to a person than her desires are. Such considerations may have been behind John Christman's claim that autonomy is the foundation of personal values: "autonomy is at the foundation of all personal values a person can call her own, since autonomy ... is nothing but the specification of that which having desires of one's own turns out to mean."[19] If we need an account of autonomy to determine which of our values are themselves authentic, truly our own, then Watson's theory of values cannot replace that of second-order desires on the bi-level view. That we need such an account of the authenticity of values as well as desires can be seen in the following example. Girls who are socialized under patriarchy often acquire and internalize values which support their oppression and subordination as women. If they experience renegade desires which conflict with their beliefs and associated values concerning "the proper role of women" and "feminine virtues", they will judge those desires to be bad and unworthy of gratification. But one could surely question whether such values are truly their own (even though they have internalized them).

Finally, Watson insists that evaluations are typically of possible courses of

---

[18] Cf. Lawrence Haworth, Autonomy: An Essay in Philosophical Psychology and Ethics (Yale University Press, 1986), pp. 55-58.

[19] John Christman, "Introduction" to The Inner Citadel: Essays on Individual Autonomy (Oxford University Press, 1989), p. 19.

action or alternative states of affairs, and so are first-order. Yet he recognizes that sometimes what one values is a particular desire or appetite. One might, for example, value one's appetites for food and sexual gratification.

> Part of what it means to value some activities in this way is this: we judge that to cease to have such appetites is to lose something of worth... The judgement is ... that it is of value to have and (having them) to indulge these appetites.[20]

This sounds indistinguishable from second-order volition-formation, for to say that one approves of having and acting on certain desires is just to say, on the bi-level view, that one identifies with them.

The core of Watson's disagreement with Frankfurt, then, must depend upon his claim that most evaluations are of possible outcomes or actions and so are first-order. This is, of course, an empirical question which philosophers are not uniquely suited to answer. But if autonomy depends upon self-evaluation in some important sense, then at least some of the time autonomous agents must be concerned with evaluating their own desires, plans, projects and values, for these are important features of their self-conception. When they are engaged in such evaluation, moreover, they may bring other considerations than moral considerations of goodness or worthiness into their practical deliberations; for they might approve of a particular plan for prudential reasons, or endorse a desire just because its gratification is expected to bring them pleasure.

The insistence that values are first-order commitments raises a further

---

[20] Watson 1975, p. 115.

difficulty, which is related to the possibility of being heteronomous with respect to one's values but differs from it. If desires and values are both first-order, with actions or states of affairs as their objects, how are conflicts between them to be resolved? Why is the presumption of authority given to values? This latter is a reasonable question to ask, especially when one recalls that a person's judgments concerning the worthiness of various alternatives may rest on beliefs which are the product of manipulation or deception, or which are indefensible on normative grounds. A racist bigot may believe, in her cool and non-self-deceived moments, that blacks are inferior to whites and so are less deserving of respect and concern; such beliefs may ground various values, such as holding membership in an all-white golf club. What superior authority do such values have over other desires she might have but not judge to be morally good, such as humanitarian desires?

Together these considerations support the conclusion that Watson has not offered a successful alternative to the bi-level account of authenticity as the foundation for personal autonomy.

I.3.ii Are Second-Order Desires Explanatorily Redundant?

The next challenge which one might raise to the suggestion that higher-order endorsement of one's reasons for action is necessary to confer authenticity upon them comes from those who argue that second-order desires are explanatorily

redundant.[21] This objection is raised by those who think that the theory of rational choice can explain the role of desires in motivation at the first-order in such a way that we can identify what an agent most wants to do without resort to a higher order. I shall argue that second-order desires play a role in motivation which cannot be reduced to one's highest-ranked first-order desire and so they are not redundant.

The challenge, then, is that talk of second-order desires is unnecessary because maximizing views of practical rationality can do all the explanatory work which second-order desires do. The first reason to think that we need recourse to higher-order desires is that we can and do form preferences about our preferences. We care not only about what we do but why we do it. This is the role which second-order volitions occupy in bi-level theories of motivation. But someone who is committed to admitting only a single comprehensive preference ordering may not agree that this requires recourse to second-order preferences. Such a person could agree that we form preferences about our preferences, i.e., that we can take our preferences over first-order outcomes to be themselves possible outcomes of choice. To do so an agent must construct a second utility measure for preferences over these as outcomes. This is possible, because one can

---

[21] This challenge has not been made clearly in the literature, though it has been suggested to me in conversation with Duncan MacIntosh and is often raised in discussions of the bi-level theory of autonomy. Something like it was urged by Irving Thalberg in "Hierarchical Analyses of Unfree Action" Canadian Journal of Philosophy (1978), though he was concerned with the more limited thesis that invoking the bi-level theory of desires presents a misleading account of coerced action, which can be better explained on a uni-level theory.

construct a utility measure over any set of objects taken as possible outcomes of choice, including one's desires or preferences themselves. I am inclined to think that if one takes one's desires as possible outcomes of choice then we are already, at least surreptitiously, invoking a second-order scale of preference.

But the challenge is deeper than this response allows, for the charge is that one's preferences over preferences are explanatorily redundant: the second-order preferences are completely parasitic on the first-order preferences over outcomes. Within versions of choice theory which take seriously the need to form preferences about our preferences, as a way of solving certain kinds of social choice problems,[22] the ranking of one's preferences over preferences is ultimately explainable and justified simply by reference to one's first-order preferences over outcomes as states of affairs which are brought about by one's choice or action. One forms preferences about one's preferences when changing one's first-order preferences would itself be a maximizing strategy for achieving those outcomes one has ranked highest before the choice of the new preference. For example, in the classic prisoner's dilemma, two prisoners face the following choice situation: they must choose whether or not to confess to their crime. What will be determined by their choice is the length of time they will spend in prison. This is an outcome over which both prisoners have preferences: they both prefer to minimize their jail time. But given the way the dilemma is constructed, if they

---

[22] David Gauthier's Morals by Agreement (Oxford University Press, 1986) is an example of the kind of theory I have in mind here.

choose on the basis of their preference to reduce their jail time to a minimum then they will both wind up spending more time incarcerated than they would if they acted on a different preference. Given this, they would achieve an outcome that they rank higher (less jail time) by changing their preferences. If they can make their preferences an object of choice then they have reason to choose a different one from what they both have. But this is because of the utility that they both accord to minimizing jail time; the change in preference can be justified because it would achieve this outcome. Thus the second-order utility measure is entirely parasitic on the first-order measure over outcomes. While the bi-level talk might be more convenient than talking about preferences over preferences as possible objects of choice, it is the first-order preference which carries the explanatory and justificatory weight at both levels.

Clearly, this differs from the role which second-order desire-formation plays within a hierarchical account of authenticity. When we engage in second-order reflection on our preferences what we may be worried about is the worthiness of our preference for some outcome. Revising our preferences so as to achieve that outcome, should that be necessary, cannot be what we are doing when we form a higher-order preference to resist some preference which we have because we judge that preference not to be worthy of satisfaction.

Preference-revision on hierarchical accounts need not be based on means-end rationality, which is the role it plays within maximizing accounts of practical rationality, where forming preferences about our preferences is required as a

means of satisfying those very preferences. It may be that I have a strongly felt desire to get revenge for a wrong which has been done to me, for example. If revenge is the outcome I most prefer (or which I believe would afford me greater pleasure than not 'getting even', to weaken the case a little), the only reason I would have for revising this preference on a means-end account would be if doing so led to a higher probability of achieving that outcome (or other outcomes which I desire before the preference change). But there is room for a different kind of evaluation of this desire. It might be that I believe I could satisfy my desire for revenge without getting caught, or without sacrificing other ends which I have. Yet I might still resist this motive as a reason for me to act.

I might have moral or religious reasons to resist this desire, for example, which would lead me upon reflection to reject this desire as one I should act on. Or I may simply be the kind of person who wants not to act on desires which are petty, or destructive, or hurtful to others. Such desires may be inconsistent with my self-conception as a tolerant or forgiving person. Likewise, in the case of the unwilling drug addict we have been considering, his higher-order preference to resist his desire for the drug does not seem to be parasitic on a preference-ordering between the conflicting first-order desires for and against taking the drug at all. He may anticipate considerable pleasure from taking the drug and real discomfort and pain from not taking it. Yet he may nonetheless conclude that he does not want to act on his desire for the drug.

The defender of the maximizing view could, at this point in our debate, claim

that these examples are not counter-examples to his uni-level position, for their plausibility presupposes that the agent has other preferences -- to do what is morally right, to obey the dictates of the church, to not be hurtful, etc. -- which are ranked higher than the desire which is rejected; indeed, the rejected desire is rejected just because its satisfaction would conflict with those higher-ranked preferences with which it is inconsistent. On the maximizing theory, then, this revision of preference, or the action of resisting the rejected desire, would be rationally explainable at the first-order.

However, this response does not allow that a desire might be rejected because one *believes* it is immoral or otherwise unworthy of gratification, even if it conflicts with no other desires the agent has. There are many normative standards which might be brought to bear in the evaluation of one's desires, though, beyond one's other preferences. Thus I allow, while the maximizing theory does not, that one could reject a desire just because one believes it is immoral, for example. The following example may illustrate this point. A man may want to have sex with a woman of his acquaintance; this desire, we can suppose, does not conflict with any other desires that he has, and he believes that gratifying it would produce much positive utility for him. But he also believes that it would be morally wrong to act on this desire. If he then withdraws from the desire as a reason for action, and successfully resists its motivational sway, he has denied himself a source of utility needlessly, and the defender of the maximizing view would claim that both the decision to alter his desires and his

subsequent abstemious actions are irrational. But surely the man in our example can claim that, at least to his own self, he was true. The decision and subsequent actions would exemplify the properties of authenticity, even though they would fail the maximizing test of practical rationality.

Whatever the force of the foregoing considerations, there is a second reason for thinking that the charge of irrelevance is not warranted as a criticism of the bi-level theory of autonomy. For maximizing accounts of rationality seem to be aimed at answering different questions from those which an agent asks when she wonders if she is self-directed in what she desires and what she does. When she is wondering if she is autonomous in her actions she is not, at least primarily, asking if her action was rationally explainable or justifiable. It may be both; yet she may feel estranged from the desire which issues in the action or be completely heteronomous in regard to it. Suppose, for example, that a benevolent neurosurgeon has manipulated her desires and beliefs in such a way that she cannot fail to act rationally, in the sense of acting efficiently relative to those desires and beliefs. Such manipulation, provided that the neurosurgeon only implants in the beneficiary (victim?) desires which are ordered, complete, transitive and suited to her external circumstances in such a way that their satisfaction is possible, need not defeat the conditions of practical rationality or rational agency. As long as her actions are caused by her desires, irrespective of the origin, agency and choice are retained. This might seem like a gift for those who worry about their rationality, but it will not be welcomed by those who are

concerned with their autonomy. Conversely, acting on some desire may satisfy the conditions for autonomous action, yet not be justifiable on rational grounds.

Even non-autonomous agents can have preferences that form an ordered set. As such, they would be capable of rational action. But theories of autonomy must be concerned with the origin of those very desires: were they adopted uncritically from others, were they hypnotically implanted, etc.? Theories of autonomy must also be concerned with the attitudes of agents towards their desires and the ranking they give to them. Whereas theories of rational choice take the preferences of persons as basic, theories of autonomy do not, and so the latter need the conceptual resources to raise questions about the relationship between preferences and those whose preferences they are. The bi-level theory provides those resources through its account of authenticity.

The thought behind this objection is not just silly, however. We might expect that our theories of autonomy and practical rationality would largely overlap; it would be very bizarre if our theory of autonomy implied that only irrational actions or agents were autonomous.[23] But it would be equally suspicious if our conception of autonomy implied that one could never be self-directed and yet act irrationally. Perhaps the proper way to understand the relation between these two theories is that the theory of autonomy should identify those desires which are autonomous and then the theory of practical rationality should tell us how to

---

[23] Cf. Lawrence Haworth, Autonomy: An Essay in Philosophical Psychology and Ethics (Yale University Press, 1986). I discuss the question of the necessity of rationality for autonomy more fully in Chapter V.

reason so as to maximize their satisfaction.

These considerations support the claim, even if they do not conclusively demonstrate, that uni-level theories of desires are inadequate for the modelling of theories of autonomy. In the next two chapters, I will present further reasons in favour of adopting the bi-level theory of desires and identification as necessary for authenticity and so for personal autonomy.

CHAPTER II

IS IDENTIFICATION NECESSARY FOR AUTONOMY?

II.1 INTRODUCTION

Is identification with a desire necessary to confer authenticity upon it? Is the creation of an authentic self necessary for self-direction? These are the questions to which we must turn if we are to appreciate the significance of the bi-level theory of autonomy offered by Dworkin, Frankfurt and Haworth. I shall argue in this chapter that identification with a motivating desire is necessary for its authenticity, and that having authentic desires is necessary for being self-directed. (I shall often speak of identification as a necessary condition for autonomy, understanding implicitly that authenticity is the concept which relates identification to autonomy.) First I shall examine Dworkin's recently developed reasons for rejecting this claim, arguing that they fail, and then I will offer more positive reasons for thinking that identification with some desires is necessary for personal autonomy.

II.2 DWORKIN'S REASONS FOR ABANDONING THE NECESSITY OF IDENTIFICATION

In his original (1976) statement of the bi-level conception of autonomy Dworkin

argued that one's higher-order endorsement of some of one's first-order desires conferred authenticity upon them in virtue of the agent's "identifying" with them. Dworkin argued that identifying in this way with the first-order desires by which one was motivated is a necessary but not a sufficient condition for being autonomous.[1] Anticipating the challenge that one's second-order desires might themselves be adopted as a result of external influences, he argued that identifications must be made under what he refers to as conditions of "procedural independence".

> A person is autonomous if he identifies with his desires, goals, and values, and such identification is not itself 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.[2]

This distinction has been notoriously difficult to make, yet there are clear paradigms of processes which subvert the critical and reflective capacities of agents, and so undermine the effect of identification with their desires (e.g., brainwashing, hypnosis, indoctrination, drugs, manipulation of information, etc.), as well as those which promote them.

Dworkin now thinks this is mistaken and has abandoned this specification of the bi-level account of desires. Although he still treats autonomy as a capacity

---

[1] Dworkin 1988, p. 15. His original argument can be found in Gerald Dworkin, "Autonomy and Behavior Control", Hastings Center Report 6 (February 1976).

[2] Dworkin 1981, p. 61.

to reflect upon and adopt attitudes towards one's first-order desires, and believes that one "defines oneself" through those attitudes, he no longer maintains that identification with one's desires at the second-order is necessary for them to be authentic or for an agent to be autonomous with respect to them. Rather, he now believes that autonomy is best understood as a capacity to engage in critical reflection on one's motives and alter them when deemed appropriate. It is the "ability both to alter one's preferences and to make them effective in one's actions,"[3] rather than any specific act of reflection and identification with one's motives, which Dworkin now sees as characteristic of autonomous agents. I shall begin my examination of the question of whether identification with a desire is necessary for its authenticity by looking at Dworkin's reasons for abandoning this condition.

Dworkin gives four arguments in support of this alteration of his doctrine. The first is the possibility that one can be heteronomous at the level of one's volitions.[4] His concern is that one's autonomy can be interfered with by processes, such as being kept ignorant of relevant information or otherwise being manipulated, which do not themselves interfere with one's identifications. This possibility implies, however, only that identification with one's desires is not sufficient for autonomy with respect to them; it does not imply that identification

---

[3] Dworkin 1988, p. 17.

[4] The problem of heteronomous second-order desires was, I believe, first introduced in Campbell 1979, pp. 148, 212-214. We shall return to at some length in Chapter IV of this work.

is not necessary.

The second reason Dworkin has for abandoning identification as a necessary condition of autonomy is that it places an undue emphasis on achieving congruence between one's higher and lower-order desires. This is the problem of the "happy slave". Briefly, the worry is that one could face severely restricted options (indeed, one could be a slave or highly constrained prisoner), yet one could still be autonomous so long as one identified with those first-order desires one was free to act on. On the bi-level view those first-order desires with which the slave identifies would be authentic. Thus, a contented slave would be autonomous while a discontented slave, who had many desires with which he identified but which he was not free to act on, would not be autonomous.[5] Like the worry considered just above, this concern speaks against identification's being sufficient but not against its being necessary. For the slave achieves whatever autonomy she does on this model by identifying with her first-order slavish desires. As Dworkin himself realizes, the problems with such revisions of preferences will have to be dealt with by specifying the constraints on procedural independence adequately, rather than by abandoning the necessity of identification.[6]

---

[5] This worry was first articulated by Isaiah Berlin, "Two Concepts of Liberty", in Four Essays on Liberty (Oxford University Press, 1969). We shall return to it in Chapters III and VII of this dissertation.

[6] My claim here is not that this is a trivial objection, but that the process of identification is not the source of its force and so abandoning the necessity of identification will not answer it.

Identification is problematic also, Dworkin says, because it allows the possibility that one can identify with a first-order desire which is irresistible, which will be effective in determining one's action and which cannot be revised simply through critical reflection upon it. So long as an agent identifies with that desire then she is autonomous with respect to it. This seems to some people to be unintuitive, for a drug addict can be autonomous on this view so long as she has identified with the addiction which moves her. Again, the problem posed by irresistible desires' being sanctioned at the higher order and so conferring authenticity upon them, if one finds this counterintuitive, is that identification is not sufficient to ensure that the person is autonomous with respect to them.

As I will not return to this topic until the final chapter of this thesis, and there only indirectly, I will deal with this argument at greater length here, thus postponing the discussion of Dworkin's final reason for abandoning the necessity of identification. My intuitions differ from Dworkin's on questions posed by irresistible desires. As a way of illustrating that my intuitions have at least some pull, consider the difference between what Frankfurt calls "willing" and "unwilling" drug addicts. In both cases, we can suppose, the addictive desire for the drug is irresistible: it will determine the actions of the addicts, no matter what they choose to do. But the willing addict may identify with his desire to take the drug; he may like the sub-culture which goes with such activities and the people with whom such behaviour brings him in contact, as well as the pleasure which he derives from the drug-taking itself. In such a case the fact that the desire is

addictive seems to make it no less internal to him than any other desire he has.

Furthermore, the irresistibility of the desire may play no role in explaining his drug-taking. Indeed, the willing addict may even be unaware that the desire for the drug is irresistible, that he is addicted to the drug. If the irresistibility of the drug is irrelevant in explaining his behaviour, then it is unclear what purchase we can get from its irresistibility alone in claiming that it must be external to him. The irresistibility of a desire undermines autonomy, surely, only if the person wants to resist its influence. Like Frankfurt, I think we want to be able to distinguish between these two addicts. The desire to take the drug is properly attributable to the willing addict, but not the unwilling addict; the willing addict acts as she most wants to when she takes the drug, while the unwilling addict does not; the willing addict may be morally responsible for taking the drug, while the unwilling addict may not be. The willing drug addict, I submit, acts autonomously when she takes the drug, while the unwilling addict does not.

It is because the willing addict has identified with the desire to take the drug, I want to say, while the unwilling addict has rejected it, that we can make these distinctions. Identification seems necessary for distinguishing between cases in which the same action is a self-directed act for one agent and not for another (or for the same agent at different times).[7] If we follow Dworkin and concentrate on

---

[7] There is an ambiguity in Dworkin's argument concerning irresistible desires which I shall just mention here. The capacity to identify with some of one's desires and reject others as reasons for acting is necessary to be an

the capacity to reflect critically upon and alter our desires as necessary for autonomy, then we cannot distinguish between the willing and unwilling addict: they are equally incapable of altering their desires should they desire to do so and are both thus lacking in autonomy.

The problems posed by irresistible desires are not challenges to the necessity of identification within bi-level theories of desires, however. Cases like that of the willing addict, who could come to approve of his addictive desire and so be autonomous with respect to its influence on his motivation, are cases in which the agent has identified with a desire but is still not autonomous (in Dworkin's view) in relation to it. Even if this is correct, such cases show only that identification is not sufficient to confer autonomy on a desire with which he identifies.

To support the claim that identification is not necessary, Dworkin would have to provide an example of an agent who was autonomous with respect to a desire but who had not identified with it. In deciding whether it is possible to construct

---

autonomous agent. But the exercise of this capacity in, say, actually rejecting a particular desire, does not ensure that it will be effective in determining what the agent does. We need to distinguish between the capacity to adopt autonomous motives through the process of identification or rejection and the actual condition of acting autonomously, i.e., acting on those desires which one has identified with and resisting those from which one has withdrawn. We need, in short, the distinction between autonomy as it is predicated of agents, and their desires and actions. I draw such a distinction in the next section of this chapter. Dworkin's new formulation of autonomy, as involving both the ability to change one's desires and act on such changes, is a formula for *de facto* autonomy, i.e., for autonomous action. The problem posed by irresistible desires is not that one might approve of them and so confer autonomy upon them, but that one cannot effectively reject them. That is, irresistible desires do not threaten the capacity to be autonomous, and they defeat de facto autonomy only if the agent wants to resist them.

such a case we must be careful to distinguish "not identifying with a desire" from "rejecting a desire": the former is merely the lack of identification as positive endorsement, while the latter is its converse and is constituted by positive withdrawal from the desire in question. If an agent merely has not reflected upon and identified with a desire, then, I want to claim, the agent is not autonomous with respect to it.[8] We do not know whether he "defines himself" in terms of it, or whether he could alter it if he chose to do so. If, for whatever reason, the agent could not reflect upon a desire and identify with it or reject its influence as a reason for action, then on either of Dworkin's conceptions of autonomy he would not be autonomous with respect to it.

The other possibility is that the agent has actively rejected a desire as a motive for acting after reflecting critically upon it. Insofar as the agent now resists the influence of that desire he does so autonomously, i.e., he is autonomous with respect to it. But this is not a case of an agent's being autonomous with respect to some desire in the absence of an act of identification: it is simply the negative instance of identification. Indeed, it is plausible to say that as a result of such an act of withdrawal the agent has also come to have and identify with another desire -- the desire to resist the first desire.

It would seem, then, that it is not possible to construct a case in which an agent is autonomous in relation to a desire which he has not identified with or withdrawn from after reflecting critically upon it as a possible reason for action.

---

[8] Cf. Section II.3.

If this is true, then Dworkin cannot succeed in claiming that identification is not necessary for autonomy.

There is an alternate reading of Dworkin's claims, however, which may allow him to escape the dichotomy just developed and with which I have no quarrel, once it is duly qualified.[9] It may be that in claiming that identification is not necessary for autonomy Dworkin means that one may have failed to reflect upon and form higher-order attitudes of approval or disapproval toward some specific desire, yet he could still be an autonomous agent, in the sense that he might have the capacity to reflect upon and alter it if he chose to, even though he has not done so. This surely is correct, if autonomy is understood to be the capacity to engage in second-order reflection and endorsement of some of one's desires. But even this does not entail that some specific acts of identification are not necessary, nor does it entail that one can be autonomous with respect to a specific desire (i.e., that the desire can be authentic), or in performing the action to which it gives rise, without identifying with it.

This reading of Dworkin is supported by the final reason he gives for abandoning the necessity of identification within his account. His fourth reason for abandoning the necessity of identification is that autonomy is a global rather than a local concept. Specific acts of identification are temporally confined and range over particular desires, yet autonomy "is a feature that evaluates a whole way of living one's life and can only be assessed over extended portions of a

---

[9] Cf. Section II.3.

person's life".[10] It is true that "being autonomous" is not something one can be only at an instant. But there is an ambiguity here. The capacity to identify with or reject the reasons which lead one to act is a capacity which one must have over extended portions of one's life if one is to be an autonomous agent. A single act of identification would not ensure that a person was autonomous in this basic sense. Yet one might have the capacity to reflect upon and form attitudes towards one's desires without having exercised that capacity on every desire which one has or which one acts on. In this sense one can be autonomous without having identified with a particular desire which issues in action. But this sense of autonomy cannot be predicated of actions; an action is not autonomous just in virtue of being the action of an autonomous agent. If the person is autonomous in what he does, then reflection and second-order assessment must have been engaged in on the specific desire which motivates him. Thus, although a person can be an autonomous agent without having identified with every desire which moves him to action, his actions are autonomous only insofar as they issue from desires with which he has identified.

If we do not draw this distinction between autonomy as it is predicated of agents and autonomy as it is predicated of their actions, then the following implausible result is licensed: an agent who has the capacity for autonomy could act on a desire which he found morally repugnant by his own standards, which he could not claim as his own and which has no place in his self-conception, and

---

[10] Dworkin 1988, p. 16.

the agent would be autonomous in so acting. What this shows is that, while one can have the capacity to be self-directed without having identified with all the desires which one acts on, one's actions must be motivated by desires which have actually been endorsed at the second order if they are to be self-directed actions.[11] Thus, identification is necessary for autonomy in action (what I call *de facto* or act autonomy). Dworkin's arguments support only the contention that identification is not sufficient for autonomy.

Now Dworkin might concede this last point, yet continue to claim that, since autonomy in its most basic sense is a capacity of agents rather than a property of actions, and identification is not necessary for that capacity, identification is not necessary for autonomy. Even this will not do, however, for the capacity to be autonomous cannot be analyzed independently of its exercise. One would not *be autonomous* simply by virtue of having the capacity to structure one's motive through acts of identification or rejection; one must exercise the capacity, i.e., one must actually identify with some desires and reject others. An autonomous

---

[11] This may seem too strong, or too intellectualist, a requirement for autonomous desires or autonomous actions, for it requires that one has actually engaged in reflection upon and endorsement of a particular desire if one is to be autonomous with respect to it. I do not mean to deny, however, that one could have very good reasons for judging that an action was autonomous even though the agent had not explicitly endorsed its motive in this way. If it falls under the general category of benevolent actions, for example, and the person has endorsed benevolent motives as reasons for action in the past, or if she performs some action repeatedly without showing any signs of regret, remorse or guilt after doing so, etc., these are good indications that the action was self-directed. But the possibility remains in such cases that the action was motivated merely by habit, or social expectations, which the agent would reject if she reflected upon them. Explicit reflection bars this possibility.

person must have the capacity to structure at least some of her desires in a way which incorporates them into her self-conception and sometimes exercise it.[12]

Furthermore, if autonomy is a feature that calls for evaluating a whole way of living one's life and can be assessed only over extended portions of a person's life, then autonomy requires not only actually identifying with some desires and rejecting others (and so identification is necessary) but also acting in conformity with the desires which have been sanctioned and resisting those which have been rejected. To be an autonomous agent, to define oneself through critical assessment of possible motives for action, requires that one form relatively persistent identifications, and that those identifications ground persistent autonomous action. These considerations, while they support Dworkin's contention that it is the capacity to reflect critically upon and adopt attitudes of approval or disapproval towards some of one's desires which is central to autonomy, do not support abandoning identification and the capacity to form authentic desires as necessary conditions of being autonomous.

II.3 AGENT AUTONOMY, DESIRE AUTONOMY AND ACT AUTONOMY

---

[12] This is so even if we drop the identification criterion itself, and adopt instead Dworkin's more recent formulation. Having the capacity to reflect critically upon one's desires and alter them when required cannot be understood in isolation from its exercise. For one must actually reflect on one's desires and form attitudes about them as possible motives of one's own if one is to "define oneself" in relation to them.

My response to Dworkin's rejection of identification with a desire as necessary for its authenticity, and hence its autonomy, relied in crucial ways upon distinguishing between autonomy as it is predicated of agents and as predicated of desires and actions. This distinction must be made clearer and defended before proceeding.

Agent autonomy is a property predicated of persons, on the bi-level view, if and only if they have the psychological (cognitive, conative and affective) capacities necessary to critically reflect upon and identify with or reject their first-order desires and beliefs as reasons for action. The defining characteristic of autonomous agents is that they have the basic capacity to reflect upon and structure their motives in a way which makes self-definition possible.

But autonomy is also predicated of actions and desires[13]. While it is necessary for an action or a desire to be autonomous that it be an action or desire of an autonomous agent, it is not sufficient to ensure that an action or desire is autonomous that it is ascribable to an autonomous agent. The reasons for this latter observation may seem too obvious to require stating explicitly, but the distinction between agent, desire and act autonomy will be important in what follows and so I shall say something more about it here.

It is a feature of autonomous agents that they have the capacity to reflect upon and endorse (or reject) certain of their desires as reasons for action. It is a

---

[13] I shall use "desire" and "motive" interchangeably, since I mean to cover with the former term any conative state which could motivate intentional action.

further feature of autonomous agents that they sometimes fail to act on those desires they have endorsed and act on those desires they have rejected. In short, autonomous agents sometimes perform actions which are not themselves autonomous. The phenomenon of weakness of will provides ready examples of autonomous agents who act non-autonomously in particular instances.

The bi-level theory of desires offers a very insightful account of weakness of will. This has been widely discussed in the literature, however, and I want here only to employ that account to illustrate how a person who has the requisite capacities for personal (agent) autonomy may, nonetheless, fail to act autonomously. Oft-discussed paradigms of weakness of will are displayed by unwilling drug-addicts or smokers. Since in both these cases, however, the desires over which one's will is weak might be thought of as addictive and so irresistible (and so they might pose independent problems for autonomy), I shall offer a different example. Let us imagine a woman who has all the cognitive, conative and affective capacities needed for autonomy. These capacities allow her to reflect critically upon her reasons for action and to make those she approves of authentic through the process of identifying with them, while she rejects others. We can also suppose that she exercises sufficient self-control that her actions reflect her second-order commitments, so that she acts on those desires she approves of and resists acting on those motives she has rejected, in most cases. But our heroine is also attracted to a married man, and when she succumbs to her desire for him she acts in a weak-willed way. She has a first-order desire to be

intimate with him (or, more probably, a cluster of first-order desires to spend time with him, to be his lover, etc.), but she also has a first-order desire to resist his attractions. She has reflected upon these desires, examining the reasons in favour of and against both sets of first-order desires. Given that she desires many other things which she (correctly) believes are incompatible with satisfying her desire to continue her intimacy with this man, she forms a second-order desire to resist her first-order desire for him. She does not want it to be effective in leading her to action. Insofar as she continues to pursue her involvement with him after she has rejected her desires for that association, she acts contrary to the desire she wants to be effective (the desire to resist his charms); she may feel alienated from her effective desires, experience regret or even shame when she continues the liaison, etc.. Though she has structured her motivations in a way that confers authenticity on one of her desires, she fails to make that structure effective. Her action, when she succumbs to the attractions of her lover, is not autonomous, though she is an autonomous agent capable of self-direction. Autonomous action requires not only the general psychological capacities which make agent autonomy possible, but the self-control[14] to carry out one's autonomous projects and act on one's authentic desires as well. When a person acts from weakness of will, she has the abilities for self-definition, but lacks the

---

[14] Haworth argues that self-control is a necessary condition for living autonomously, in Haworth 1986, Chapter 2.

specific control needed to act autonomously.[15] Her action is not an instance of self-direction, and does not express her authentic self.

A different case of weakness of will might be displayed, more briefly, by an aspiring dancer. Imagine a woman who desires very much to be a *prima ballerina*, and who identifies wholeheartedly with this goal after subjecting it to critical scrutiny. She knows, furthermore, that strenuous practice, continuous dieting and abstinence from certain practices (drinking alcohol, staying out late, etc.) are required if she is to attain her goal, and so she desires to resist her desires for such things as instrumental to her plan of becoming a successful professional dancer. Yet when invited to parties she accepts, she drinks too heavily, she does not exercise and practice enough, etc.. When she engages in these activities, she acts from weakness of will. She has formed a second-order volition to become a dancer, and to do whatever it takes to accomplish that end, but she is unsuccessful in making her volition effective in determining her actions.

The features which make the behaviour of our unwilling adulteress and our would-be ballerina instances of weakness of will can be generalized so as to give the following characterization of weakness of will: A person P is weak-willed

---

[15] If a person is prone to certain sorts of failure of control, this may indicate that her capacity for autonomous desire-formation is, itself, impaired with respect to those desires against which she acts in a weak-willed way. It may signal a lack of wholeheartedness in her identification, or self-deception. It is true, nonetheless, that one could display all the capacities necessary for agent autonomy, with respect to a wide variety of other desires, plans and projects, yet fail to act autonomously due to weakness of will with respect to a particular desire.

relative to some first-order desire D[16] if and only if 1) P has reflected upon D and endorsed it as a reason for action, making it the object of a second-order volition V (to act on D in suitable circumstances), yet acts contrary to D in circumstances which make the satisfaction of D possible for P; or 2) P has reflected upon D and rejected it as a reason for action, making it the object of a second-order volition V′ (to not act on D), yet acts on D in circumstances in which P could have resisted D's motivational influence. Given this characterization of weakness of will we can see that to be weak-willed one must have the capacities of critical reflection and volition-formation and one must act contrary to one's second-order volition, either by acting on a desire which has been rejected or by failing to act on a desire which has been endorsed. Hence, all weak-willed actions will be actions of autonomous agents but they will fail to be autonomous actions.

Autonomous actions, by contrast, are those of autonomous agents and are motivated by desires which have been approved of at the second order. (Again, this must be qualified by the conditions which are required to ensure that identification with a desire is not just necessary to confer autonomy upon it, but also sufficient.) The distinction between agent autonomy and act autonomy is, then, a distinction between being able to structure one's motives by reflecting critically upon them and making some of them authentic reasons for action, on

---

[16] In characterizing weakness of will as necessarily relative to some specific desire I am, here, contrasting the phenomenon with the general "character trait" of weakness of will. For an interesting discussion of weakness of will as an (immoral) character trait, see Thomas Hill Jr., "Weakness of Will and Character", Philosophical Topics Vol. XIV, No. 2 (Fall 1986).

the one hand, and having those desires one has endorsed be effective in motivating one's action (or resisting those which have been rejected), on the other. Act autonomy is the actual condition of directing oneself.

One way of expressing this point might be to say that self-direction requires the ability to form autonomous (authentic) desires and the self-control to act in accordance with such motives. This would require that we also understand what it is to have "autonomous desires". This latter notion might be given a weaker or stronger characterization. A weak understanding of "autonomous desires" would be given by someone who insisted that reasons for action are autonomous just in case they have not been rejected by the autonomous agent whose motives they are. This would not require that one actually has evaluated and endorsed a desire for it to be autonomous. There may seem, on first glance, to be good reasons to adopt this approach, for presumably autonomous agents do often act on desires which they have not reflected upon or tested against their self-conception. This is probably a good thing, too, given the scarcity of resources (including time) which persons have. Any conception of autonomy which required that agents reflect on all of their desires, or even all of their desires which move them to act, would make it impossible for agents to be both autonomous and to lead very interesting lives at all.

These considerations do not support adopting a weak conception of autonomous desires, however. They do support adopting a characterization of agent autonomy which depends upon the capacity to form authentic motives

rather than upon an insistence that to be an autonomous agent one must have actually evaluated all of one's effective motives. The weak conception of autonomous desires seems to rest on a certain presumption, namely, that the effective reasons for action which motivate autonomous agents are themselves autonomous (or at least authentic) and that this presumption holds unless it is countered with specific evidence that the agent has rejected the motive. There is no good reason to accept such a presumption. Indeed, given the insights of Freudian psychological theory and the force of habit in determining human behaviour, I am inclined to reject it at the theoretical level.[17]

I would urge, then, that we adopt a strong criterion for "autonomous desires", i.e., that a desire is autonomous only if it has been reflected upon and endorsed as a reason for action by the autonomous person whose desire it is (i.e., only if it is authentic).

Regardless of how one characterizes autonomy with respect to desires or motives, however, the general distinction between agent autonomy and desire autonomy is central to bi-level theories of autonomy. It is worth mentioning, too, that it is central to non-hierarchical accounts of autonomy as well (Richard Brandt's, Gary Watson's, John Christman's, Susan Wolf's, etc.); for it must be drawn in any theory which characterizes autonomy in terms of some special

---

[17] This is not to deny, of course, that there may be very good political reasons for adopting a policy that consorts with this presumption when questions arise concerning the justifiability of paternalistic interferences with the choices of autonomous agents.

subset of the person's motives (those which would survive cognitive psychotherapy, those which are based upon judgments of the value of the desire or its object, those which would survive full disclosure of their etiology, those which are sane, etc.). In all these theories we must distinguish between the agent's "real" or "rational" or "authentic" motives and her more "superficial", "irrational" or "external" desires. Yet none of these theories would claim that agents are autonomous only if they have none but motives of the approved sort, or even that to be an autonomous agent one must act only on autonomous motives so characterized.

These considerations lead to some interesting contrasts between agent autonomy, on the one hand, and act and desire autonomy on the other. Agent autonomy seems to be, as Dworkin has noted, "a global concept". It consists of capacities, skills and competencies which can be assessed only over extended portions of time, relative to many desires, beliefs, values and other motivational states. Act autonomy is both temporally confined and relative to specific desires. Desire autonomy is confined to a single reason for action, and though some instances of autonomy in this respect must be fairly persistent, this is compatible with abrupt reversals in the assessment of specific motives. Both act and desire autonomy can be assessed episodically; agent autonomy seems less amenable to such an episodic approach.[18] Because agent autonomy is defined in terms of the

---

[18] I borrow the term "episodic" from Diana Meyers. Meyers offers an interesting discussion of the differences in emphasis one is led to depending on whether one focuses on a global or episodic conception of autonomy. See her "Personal

attainment and exercise of certain cognitive, conative and affective capacities, moreover, there can be degrees of agent autonomy. Those who have more refined skills, more fully developed capacities for practical reasoning or self-knowledge may be more autonomous than those with less fully developed capacities. If one takes agent autonomy to depend upon act autonomy or desire autonomy, then one is forced to the same conclusion: agent autonomy is a matter of degree. Those who act in accordance with more of their authentic desires and who exercise self-direction in the more important areas of their lives are more autonomous than those who act autonomously in fewer of their actions or with respect to fewer areas of their lives which are important to them. Both act and desire autonomy, by contrast, are not had by degrees. An action is autonomous or it is not; a desire is authentic or it is not. To say of any two actions, both of which were motivated by desires which the agent had reflected upon and approved of as reasons for action, that one was "more autonomous" or more an act of self-direction than the other makes no sense. So long as actions are motivated by desires with which one identifies[19] (or by desires which would survive cognitive psychotherapy, are valued, would survive disclosure of their etiology, or are sane), then such actions are equally acts of self-direction. The

---

Autonomy and the Paradox of Feminine Socialization", Journal of Philosophy Vol. LXXXIV, NO. 11 (Nov. 1987). See also Meyers 1989.

[19] Provided that the conditions which make identification with a desire sufficient for its autonomy (and not just its authenticity) obtain, of course.

same seems to hold concerning the autonomy of the desires themselves.[20]

II.4 WHY SECOND-ORDER DESIRE FORMATION IS NECESSARY FOR AUTONOMY

Thus far I have only argued against certain suggestions that identification with one's first-order desires at the second order is not necessary to be autonomous with respect to them. Are there not more positive reasons for thinking that identification is necessary? I think there are, and shall say something briefly about them here. Identification seems necessary in order to make sense of the distinction, first drawn by Aristotle, between desires which are internal to an agent and those which are external, given that there is a very obvious sense in which all of an agent's desires are equally "hers" and no one else's. For just as a movement can be an element in the history of my body without being an action of mine, so can a desire be an element in my mental history without its being a

---

[20] If one allows that one's second-order volitions might, themselves, need to be ranked according to some preference ordering, then one might say that a person acts "less autonomously" when she acts on a lower-ranked volition than a higher-ranked one, when both are available. I do not mean my remarks to exclude this possibility, though I find this description of the phenomenon somewhat forced. It seems better to me to simply speak of priorities within the set of autonomous desires. For more detailed discussion of priorities at the second-order see Campbell's discussion of False Priorities, in Campbell 1979, as well as Wright Neely, "Freedom and Desire", Philosophical Review 83 (1974).

desire of mine.[21] I make a desire truly mine (i.e., authentic) by identifying with it, and am self-directed when that desire is the reason by which I was motivated to act. Thus if having authentic desires is necessary for autonomy, then identification is necessary for autonomy.

Attention to the acts of identification is useful in helping us to understand some desires as reasons for action which have this special status -- they are *the agent's reasons* for actions, those she recognizes as reasons for herself. It is through the process of identification that desires become reasons in this strong sense. The principal reason for keeping in mind the kind of intimate connection which exists between an autonomous agent and the reasons for which she acts concerns the kinds of reasons which are adequate to explain autonomous actions *as autonomous*. Autonomous actions must be explainable by motives the agent identifies with if they are to instantiate the capacity for being self-directed. The claim is that, in order for an action to be autonomous, it is both necessary and sufficient that the autonomous motives of the agent explain the action.

The sufficiency of citing an agent's reasons for action will rule out the relevance of certain cases of over-determination in explaining why the autonomous agent acted as he did. If an autonomous agent can explain his own behaviour by reference to motives with which he identifies, then it is irrelevant

---

[21] I am drawing heavily here on Frankfurt 1977, especially p. 61. For an interesting criticism of this analogy and more detailed discussion of Frankfurt's account of externality, see Terence Penelhum, "Human Nature and External Desires", The Monist Vol. 62, No. 3 (July 1979), especially pp. 305-306.

to deciding whether the action was autonomous (or whether he was morally responsible for it) to cite causes which would have determined him to perform the action even if he had not formed the intention to do it. An example may help to make this clearer. Imagine that a person has joined a terrorist organization. As part of his initiation into the group, he must plant a bomb in the World Trade Building. The leader of the group is not fully convinced of the new member's loyalty, however; or perhaps he fears that the latter will lose his nerve at the last minute. To ensure that the bomb is set, then, he plants a post-hypnotic suggestion in the new member to set the bomb, which can be triggered by a signal from the leader if need be. (Suppose that this is done without the knowledge of our budding terrorist.) When the day comes, though, the new member does not lose his nerve. He has firmly identified with the goals of the organization and with the particular plan to set the bomb, and he has sufficient self-control to carry out the plan. He sets the bomb. While it is true that he would have done so even if he had changed his mind, or had formed the intention not to plant it, because the leader would have triggered the post-hypnotic suggestion, these facts are irrelevant to explaining his action. So long as citing his own motives is sufficient to explain the intention which he forms, we can conclude that in carrying out the action he intends he is self-directed.[22]

---

[22] See Frankfurt's discussion of the relevance of over-determination to ascriptions of moral responsibility for intentional actions in "Alternate Possibilities and Moral Responsibility", Journal of Philosophy Vol. LXVI, No. 23 (1969); reprinted in The Importance of What We Care About. Citations of Frankfurt 1969 are from the latter collection.

Furthermore, citing the agent's own suitably structured motives must be necessary to explain an autonomous action. To give an adequate explanation of an action which cited only physical laws or behavioral regularities, for example, would be insufficient for showing that the action was autonomous, or even that it was the action of an autonomous agent. To say that an action was an act of self-determination is to say that reference to the self must be made in explaining by what it was determined. Again, an example of over-determination may help to make this clear. A willing drug- addict, for example, will succumb to his addictive desire to take the drug, regardless of the second-order desires he has. Hence, his action could be explained simply by reference to his addictive first-order desire. But this would not enable us to say that he took the drug *willingly*, and that is the component which is relevant to assessing the autonomy of the action. An explanation which cites only one's first-order desires can explain an action as intentional, but not as autonomous.

Finally, the process of identification allows us to explain the role of self-definition in the attainment of personal autonomy. This is needed because we must be able to distinguish the "real" or "authentic" self from which autonomous actions flow and the more superficial self. Theories of autonomy must rely upon a view that there is some identifiable "real" self whose desires are truly the agent's own and whose task it is to control and direct the more superficial self of desires, impulses and passions which it might contingently happen to have at any given moment. We cannot, obviously, locate that true self simply by examining the

various reasons for action which the agent might have at the first-order. But we must also avoid suspicious metaphysical theories of the real self. Hierarchical accounts of autonomy have been said to require a metaphysically suspect and psychologically fragmented view of the self. Isaiah Berlin has disparagingly claimed that they depend upon the existence of an "inner citadel" within agents where the "true self" lives.[23] Such criticisms have some plausibility given that hierarchical accounts of desires are employed to distinguish between those desires which belong to the agent from those which do not, and given that in an obvious sense all desires an agent has are equally his own. Making this distinction, then, does seem to depend on being able to distinguish the real self as the locus of the agent's true desires from merely accidental or heteronomous motives which the agent might have. And this true self is not only claimed to be identifiable, but it is suggested that it has ascendancy within autonomous agents. Agents act autonomously insofar as their actions are directed by their true self, or issue from their real desires: the real self must control, in some sense, the accidental or superficial self.[24]

---

[23] Berlin 1969, p. 135.

[24] This objection has been raised by feminist writers, and Susan Wolf discusses it as a real worry in "Sanity and the Metaphysics of Responsibility", Responsibility, Character and the Emotions, ed. Ferdinand Schoeman, (Cambridge University Press, 1988). A recent criticism of hierarchical accounts along these lines is also offered by Marilyn Friedman, "Autonomy and the Split-Level Self", Southern Journal of Philosophy Vol. XXIV, No. 1 (1986); for a response see John Christman, "Autonomy: A Defense of the Split-Level Self", Southern Journal of Philosophy Vol. XXV, No. 3 (1987).

Bi-level theories of desires do seem committed to such a view. But we must ask whether that itself is objectionable. For, surely, it would only be objectionable if we supposed that such a self is created from nowhere, or is a transcendental reality or some such thing, and bi-level theories are not committed to such views. There is no pre-existing authentic self which the agent discovers, thereby achieving autonomy. On the bi-level theory of desires an authentic self is created through a process of self-definition. Although both self-discovery (in the sense of self-knowledge) and self-creation are involved in self-definition, neither of these involves suspect ontologies. Self-discovery requires only that one be able to reflect upon one's own beliefs and desires, that one not be ignorant of or self-deceived about one's actual reasons for action (one's desires, values, beliefs and the like)[25]; self-creation requires that one develop a self-conception that one can respect and that includes desires, values and goals which one identifies with after critically reflecting upon them. Those reasons for action which the agent endorses become elements in the self she is defining. It seems to be a strength of the theory, rather than a weakness, that it can explain through relatively

---

[25] This may seem, in a post-Freudian age, to make the attainment of autonomy impossible, given the important role which unconscious desires play in motivation and the sophisticated mechanisms by which human beings are able to deceive themselves about their reasons for action. Such desires are, *ex hypothesi*, not such that agents can make them objects of critical evaluation. Even if we grant the influence of such mental elements, however, I think that they can come under reflective evaluation through attention to the effects of their operations, as when they produce felt intrapersonal dissonance, shame and regret. Internalized cultural imperatives and the more mundane effects of socialization may pose more of a threat to genuine self-knowledge concerning one's reasons for action.

uncontroversial psychological processes how the authentic self can develop.

On the other hand, one might worry that the view of the self offered within such accounts is needlessly fragmented or conflicted. On this view persons are (or, at least, begin as) bundles of overtly or potentially conflicting desires, beliefs and values. The true self develops through a process of conflict-resolution. It is often as a result of coming to be aware of manifestly incompatible beliefs and desires that agents must make a decisive commitment of the kind which only autonomous agents are able to make. This kind of conflict-resolution should not be raised as a challenge to hierarchical accounts of autonomy, however; for in recognizing and resolving such crises persons experience moral growth, further their self-knowledge and integrate their experiences into a coherent whole. This is, then, a highly valuable feature of persons. If bi-level accounts of desires draw on it in identifying those processes by which persons become autonomous or maintain their autonomy, then they are drawing on important aspects of our capacities as moral agents.

It is a contingent truth that persons frequently are motivated in the first instance to adopt a reflexive stance toward their desires or beliefs as a result of becoming aware of an internal conflict or inconsistency, but such conflicts need not be the only motive for critical reflection. A person who was entirely unconflicted could nonetheless be autonomous, provided that she knew what her motives were and endorsed them as her own. Conflict plays as important a role as it does, not because it is necessary for achieving autonomy, but because it

provides a clear challenge to paradigms of being self-directed. The truly conflicted self, who knows not what she wants to do, is also the non-autonomous self, who cannot direct herself as she wishes. The conflicted self has no unambiguous self-conception upon which to draw in this matter. She must create that in resolving the conflict.

CHAPTER III

AUTONOMY AND PERSONAL INTEGRATION

III.1 INTRODUCTION

The observations with which the previous chapter ended suggest an important reason for adopting the bi-level theory of desires as a model for the attainment of personal autonomy: it provides a mechanism for resolving intrapersonal conflicts, through the processes of identifying with some desires and rejecting others as reasons for action. It will be argued here that the ability to adopt second-order commitments provides a unique and necessary procedure for resolving some kinds of intrapersonal conflicts, thus allowing the development of a coherent authentic self. Insofar as personal autonomy requires that there be a coherent self which directs the more superficial self, this will provide additional support for the claim that only a bi-level theory will be adequate for explicating personal autonomy. It will also be argued that the attainment of intrapersonal coherence is only a necessary, but not sufficient, condition for autonomy.

III.2 THE BI-LEVEL THEORY AS A COHERENCE MODEL

The bi-level theory of desires, understood as a coherence model, introduces a hierarchy of different levels of desires to illustrate how certain conflicts of desires

are resolved by an agent in such a way that she commits herself to one of the conflicting pair (through identifying with it) and rejects the other (through withdrawal from it). In this way the person participates in the conflict, rather than experiencing it as a passive bystander, and comes to define herself in terms of one of the conflicting impulses.

To understand the bi-level theory in this way, we must see that it offers a different kind of resolution to intrapersonal conflict from that offered by standard preference-orderings. Some conflicts of desires, where the desires are both first-order desires and it is contingently impossible to satisfy them both, are most simply resolved by ordering or ranking the desires on a single comprehensive scale of preference. From our preferences over outcomes (the states of affairs which would be brought about by our actions, say) we can construct a cardinal measure of utility on them. This ranking of outcomes according to preference is typically thought of as providing a first-order decision procedure: having ranked one's options according to preference, one ought to act on that preference the satisfaction of which has the highest utility (if one is a utility-maximizer). If one has, on a particular evening, a desire to see a play and a desire to attend a concert and one cannot do both, it is a perfectly adequate strategy for resolving the conflict to decide which one wants to do more strongly (which outcome one most prefers) and to rank it higher than the other. If one is truly indifferent between them, then one might adopt some lottery for deciding between them, such as flipping a coin. What is important to notice about such strategies, though, is that

they rank the competing desires on a single scale of preference. If one discovers after ranking the desire to go to the play higher than the desire to go to the concert that tickets for the play are sold out, it would be rational to try then to hear the concert. If one cannot satisfy one's highest-ranked preference, then one ought to seek satisfaction of one's second choice. Because the person really wants to do both these things, there is no reason to suppose that he is less than self-directed when he sees the play or, if that is impossible, opts instead for the concert.

There are other kinds of conflicts, though, which cannot be resolved simply by ranking the contending desires. Such conflicts involve desires which are more than just contingently mutually unsatisfiable: they involve non-contingently conflicting desires. If one desires to x and desires to not-x, then one's desires are formally inconsistent. Resolving this kind of conflict often cannot be accomplished simply by ranking the contending desires on a single scale of preference. It may require resort to a higher order, where one commits oneself to one and rejects the other. Even if both these desires are felt upon their first appearance to be equally internal[1] to the agent, through identifying with one and rejecting the other, she makes the latter external to her. So, for example, if one wants to live a life of celibacy and also wants to engage in sexual activity, then one could resolve this conflict by identifying with the former and rejecting the

---

[1] Here "internality" must be understood phenomenologically: it implies only that one experiences the motivational pull of a desire without having identified with it.

latter. Having done this, if one is faced with difficulty carrying out one's choice of living celibately, it would not be rational to seek the satisfaction of one's sexual desire as the "next-best" alternative.[2] This does not mean merely that one assigns one's sexual desires a lower ranking on one's scale of preferences; rather, one must reject them as candidates for satisfaction altogether.

In the case of formally inconsistent desires, so long as the conflict between them is unresolved, nothing that the agent does will satisfy her desires; necessarily, satisfying the desire to x will mean frustrating the desire to not-x. One way to resolve such conflicts, though, is to take a decisive stand in favour of one of the options and against the other. This is the strategy taken by Frankfurt's unwilling addict. It will be remembered that this addict has two conflicting first-order desires: the desire to take the drug and the desire not to take the drug to which he is addicted. While this conflict remains unresolved, it can be said that the addict wants both to take the drug and to resist taking the drug. The addict is also able to take an evaluative stance toward his desires, though. If such an evaluation led him merely to rank the inconsistent desires on a single scale of preference, he might decide that it is preferable, all things considered, to resist his desire for the drug. But if this desire could not be satisfied, due to the volitional strength of the addictive desire for the drug or for

---

[2] Frankfurt makes a similar division concerning different types of conflicts between desires, and also argues that some conflicts require resort to a different order of desires rather than ranking within a single order to solve. Cf. Frankfurt 1977, pp. 66-67.

other reasons, then he should at least feel some satisfaction in satisfying his lower-ranked preference for the drug.

This is not, of course, how many addicts feel about their addictions, and it is inadequate for explaining the full sense of *unwillingness* with which the addict succumbs to his desire for the drug. What the addict of Frankfurt's example does at the evaluative stage is to make a decisive commitment toward his desire not to take the drug and to reject his desire for it. This is a decisive procedure for resolving the conflict. The best strategy for resolving conflicts between formally inconsistent desires is often of this form; the person must decide which of the desires she wants to be effective and which she wants not to lead to action. She must, that is, adopt a second-order volition in favour of one of the contenders.[3]

In this way, adopting second-order volitions reduces intrapersonal dissonance: the addict can now say what he really wants to do. He has made the desire for the drug external to him, and if it determines his action in the future he can say truly that he was motivated by forces which are alien to him and which do not express his authentic self.

Seen in this light, reflection and the formation of second-order volitions are strategies for intrapersonal conflict-resolution, identification is a means of achieving coherence and personal integration. Through the act of identifying with some desires and withdrawing from others, we define an increasingly coherent self out of the various and competing desires which we have unreflectively

---

[3] Cf. Frankfurt 1971, p. 18.

accepted from others and from our society at large. This is, at least, the hope, though it may be something of a fiction: it is possible that things will not go well, and that one will identify with inconsistent desires. Nonetheless, through such reflective examination of their motives and beliefs agents come to endorse some of their pre-reflective desires and beliefs and to reject others; those they endorse become elements in their self-in-progress -- the self they are defining.

The original motive for such reflective activity is often to be found in felt intrapersonal dissonance. The experience of conflicting desires, feelings of regret or disappointment even though one's desires are satisfied and feelings of shame all signal a need for critical self-evaluation among those who have the capacities for personal autonomy, as well as those who are developing them. Such felt "crises", whether they be mild or severe, transient or chronic, often prompt one to introspection. Thus conflicts can be a powerful motive to engage in reflection and so they are important in a psychological explanation of how persons develop the capacity for reflection. And, insofar as they are resolved by the formation of second-order volitions, resolving conflicts is an important component in the development of a coherent authentic self. There is yet another reason to be concerned with the resolution of conflicts between one's desires, for unresolved conflicts defeat desire and act autonomy. Insofar as a person is conflicted, she is not autonomous with respect to her conflicting desires, and so she is incapable of acting autonomously with respect to them.

III.3 AMBIVALENCE

Taking the bi-level theory of desires as a coherence theory is attractive, for there are compelling reasons to think that intrapersonal integration and coherence are necessary for personal autonomy. Anomie, severe schizophrenia and related mental disorders, which produce, or are characterized by, dramatically fragmented or conflicted "selves", may place the attainment of agent autonomy beyond the reach of those who suffer from them. This has led many to observe that to be self-directed there must be a coherent self which does the directing. The processes of identification and withdrawal allow us to explain how such a self is constructed. Of course, not all conflicts are as wide-spread as in these cases. Many conflicts are more specific, affecting only a limited number of one's desires. Such conflicts need not defeat agent autonomy; indeed, it would be implausible to suppose that any real agent has thoroughly consistent desires, just as it would be implausible to suppose that anyone has completely consistent beliefs. If our theory of autonomy is not going to make that condition impossible for actual agents to attain, then this sort of global coherence cannot be a necessary condition of agent autonomy. Yet there is also a widely shared intuition that unresolved conflicts defeat desire and act autonomy. The bi-level theory provides a perspicuous way of explaining this intuition. To make this clear we must again consider the ways in which desires can be in conflict. For the intuition which drives this discussion is that a person cannot be self-directed if she suffers from

unresolved conflicts among her desires; she cannot act as she most wants to act if she does not unequivocally know what she wants. Surely such an intuition has great initial plausibility.

The attitudes which an agent can adopt toward any of her first-order desires as reasons for action are limited to three: (i) she may be indifferent toward it, having no second-order desire concerning its effectiveness; (ii) she may desire that it be effective; (iii) she may desire not to act on it. In none of these cases is there a conflict between the volition and the desire which is its object. Motivational conflicts cannot be between desires of different orders, therefore; conflicts must be between desires at the same level.

This is merely a technical point about the possibility of conflicts between desires. There can be no inconsistency between desires at different orders, because they are directed to different kinds of objects (desires on the one hand, actions on the other). It is noteworthy because it allows us to distinguish ambivalence (conflicting desires on a single level) from weakness of the will (the overpowering of a higher-order desire by a lower-order one). There is another, familiar, sense in which one can be motivationally conflicted across levels of desires, which is tied to the possibility of weakness of the will: suppose that a person (a recovering alcoholic, perhaps) wants to drink and also that she wants not to have her desire to drink be effective, but despite this second-order desire her first-order desire to drink is effective. This could be called a motivational conflict across levels. This is a case of weakness of will. The point being made

here is that this should be seen as a conflict between what the person most wants and what she does. Such conflict does not make it impossible for the person herself to know what it is that she most wants. The kinds of conflicts I am concerned with here are those which, while they are unresolved, make it impossible to determine what the person most wants, because there is no unequivocal answer to that question. These sorts of conflict must occur within a single level of desires.

An alternative way of making this point would be to say that when the alcoholic unwillingly drinks, succumbing to the first-order desire that she has rejected at the second-order, she manifests a lack of self-*control*. She has resolved her intrapersonal conflict of desires, and so defined herself with respect to her desires for alcohol, but she has failed to make that resolution effective. When a person has unresolved conflicting desires at the same order, however, she knows not what she really wants and so there is no possibility of either having or losing self-control in doing what she most wants to do (relative to the competing desires only, of course). Nothing she could do would manifest *self*-control in this case, for she has not committed herself to one or the other of the competing desires. Thus I shall concentrate only on conflicts within a single order in what follows.

We have already seen that there are two possible ways in which first-order desires can conflict. When the desires are only contingently in conflict, that is to say, they have different actions or outcomes as their objects but are not mutually satisfiable, we might say that they "compete" with one another. Until the

competition is resolved we cannot say with any confidence what it is that the person really wants to do. If autonomy consists in doing what one really wants to do then this person cannot act autonomously until she has resolved the conflict. Of course, this poses no great difficulty, for all that she must do is determine which of the competing op[illegible]ould afford the most pleasure (or is supported by the best, all thinks considered, reasons) and rank them accordingly. Such a ranking will resolve the conflict and so will allow us to say what the agent most wants to do. According to the utility-maximizing account of practical rationality, rational action will then reflect that ranking: the person acts rationally just in case she acts on the highest-ranked alternative available to her. We cannot say, just on the information provided in this example, that the action is autonomous, however, or even that the agent is autonomous. For we do not know whether she would identify with her desires so ordered were she to reflect upon them, or even that she has the capacities which are necessary for agent autonomy. Thus, while we can conclude that her desires are not autonomous while she remains ambivalent between them, we cannot conclude that they are autonomous just because she has resolved the conflict, if the resolution is effected by ranking the desires at the first-order.

We have also seen that first-order desires can be in conflict [illegible]hen they are formally inconsistent, [illegible]hat is, when one both wants to perform a certain action (or desires a certain outcome be brought about) and wants not to perform that very same action (or desires that that outcome not be brought about). Here there is a

conflict between the desire to x and the desire to not-x. Such conflicts, if unresolved, pose a serious problem for autonomy. The objects of the desires (the outcomes desired or the actions which one wishes to engage in) are inconsistent; whatever the agent does she will be frustrating one of the conflicting pair of desires.

I have argued in the preceding section that, while one might be able to resolve some conflicts of this type by ranking the desires on a single scale, it is often the case that such conflicts get resolved through the process of reflecting upon the desires themselves and adopting one as one's own while rejecting the other. When faced with formally inconsistent desires one must take a decisive stand, identify with one of the contenders and withdraw from the other. Before the person resolves such conflicts we cannot say, with respect to x, whether she wants to do it or not. It is indeterminate what she really wants to do, and so nothing she can do will count as expressing her true desires (again, relative to the conflicting desires only).

Both kinds of conflict between first-order desires described here produce a kind of ambivalence -- "volitional ambivalence", we might call it. So long as volitional ambivalence occurs there is no action which the agent really wants to do and so no action which can be described as being motivated by what the agent really wants to do. If authentic reasons for action are those which express what the person really wants, then such a person has no authentic reason for acting with respect to her conflicted desires. If self-direction consists of actions which

are motivated by authentic reasons for action, then unresolved ambivalence must make act autonomy, with respect to these desires, impossible. It seems, moreover, that adopting a reflective attitude toward one's desires and identifying with some, withdrawing from others, is necessary to resolve at least some such cases of ambivalence. For sometimes, taking a decisive stand toward formally inconsistent desires requires that one exclude one of the conflicting pair as a candidate for satisfaction altogether. The recovering alcoholic, for example, does not simply rank his desire to drink lower than his desire not to drink; he rejects the desire to drink as providing any reason for him to act at all. This is necessary for him to define himself with respect to his drinking. The person he wants to be is a person who does not want to drink, and does not, in fact, drink. In attempting to become the person he wants to be, he has to reject his desire for alcohol completely. Ranking it, even very low, on a single scale of preference, requires that he continue to accept it as a possible option, but that is precisely what he must not do. If this is correct, then reflection and identification are necessary for autonomy. That the bi-level theory of desires provides a mechanism for resolving such conflicts seems, then, to be a strength of the theory.

But by introducing a second order of desires, the bi-level theory also makes possible a more serious kind of conflict and corresponding ambivalence: conflict at the level of volitions.[4] Conflicts between volitions may involve volitions which

---

[4] I am enormously indebted in this discussion to Campbell's work on ambivalence in Self-Love and Self-Respect. Campbell was, I believe, the first to suggest that ambivalence at the level of second-order desires presents a serious

have objects which are mutually unsatisfiable for contingent reasons. A person may have many first-order desires which are contingently in competition, yet identify with them. If that is so, then ranking the contending first-order desires may resolve the competition, or a ranking of the second-order desires may be needed. Yet this poses no unique problems, and the resolution techniques for ranking alternatives within a single scale seems adequate to both tasks.

Second-order volitions can conflict in another way, by being themselves inconsistent. Inconsistent volitions are a pair of second-order desires with the same object. An agent who has inconsistent volitions has a second-order desire to be moved by the desire to x and a second-order desire not to be moved by the desire to x. Stated in terms of attitudes of approval and disapproval, the agent both approves and disapproves of the desire to x. For example, a person who both approves of his desire to exercise (because he believes that it is necessary to maintain his health, say) and disapproves of this same desire (because he believes that it has been caused by the multi-billion dollar advertising campaigns of multi-national manufactures of fitness equipment, and he dislikes the idea of being a dupe to the techniques of the persuasion industry) has conflicting second-order commitments. Insofar as they take the form of volitions, he wants both to act on and not to act on his desire to exercise. Nothing the agent can do while so conflicted will express what he really wants to do. This conflict cannot be

---

obstacle to the attainment of personal autonomy on the bi-level approach (or, as he puts it, to the pursuit of one's personal good, understood as the effective gratification of one's self-identifying desires).

resolved by forming another second-order desire concerning the effectiveness of x, for this would involve the mere duplication of a second-order volition which the agent already has. Nor will moving to a higher order of evaluation resolve the conflict; for if it is true that the agent approves of the desire to x and disapproves of the desire to x, then reiterating the approval or disapproval at a yet higher level will not resolve the conflict. When one has formally inconsistent second-order volitions one's self is in conflict.[5] Thus we might call this "personal ambivalence". The agent has failed to decisively identify with or withdraw from the desire to x. Such confiicts make self-direction impossible, for there is no coherent self to do the directing (again, relative to the conflicting desires only). No action will be expressive of what he really wants to do.[6]

Frankfurt is well aware of the threat that unresolved intrapersonal conflicts pose to autonomy. Thus he insists that a desire is authentic only if one's identification with it is "decisive" and "wholehearted"; by which he means that one's identification with it is made in the absence of any present or anticipated conflict with respect to that desire. One cannot be simultaneously ambivalent toward and wholeheartedly identified with a particular desire. Thus Frankfurt describes ambivalence as "a lack of coherence within the realm of the person's higher-order volitions themselves" and claims that this is a question "of whether

---

[5] Campbell 1979 makes precisely this point, p. 172.

[6] Though the structure of the conflict is different, the same points can be made if one identifies at the second-order with the desire to x and identifies with the desire to not-x.

the highest-order preferences concerning some volitional issue are wholehearted".

> It has to do with the possibility that there is no unequivocal answer to the question of what the person really wants, even though his desires do form a complex and extensive hierarchical structure. There might be no unequivocal answer, because the person is ambivalent with respect to the object he comes closest to really wanting: In others words, because, with respect to that object, he is drawn not only toward it but away from it too. Or there might be no unequivocal answer because the person's preferences concerning what he wants are not fully integrated, so that there is some inconsistency or conflict (perhaps not yet manifest) among them.[7]

Frankfurt's position seems to be, moreover, that in the absence of such conflict one is assured not only of coherence at the level of one's second-order volitions but also of the autonomy (and not merely the authenticity) of the desire so identified with. Thus Frankfurt takes wholehearted identification with a desire to be not just a necessary condition of its autonomy, but a sufficient condition as well.

III.4 WHY COHERENCE IS NOT SUFFICIENT FOR AUTONOMY

I have been taking it for granted that, in the absence of conflicts among one's desires, identification with a desire is both necessary and sufficient for its authenticity. But now we must ask whether identification with a desire at the

---

[7] Frankfurt 1987, p. 165.

second-order is sufficient to confer autonomy upon it provided that the identification is decisive and wholehearted (provided, that is, that the person feels and anticipates no conflict between the desire identified with and others with which she also identifies). The answer, alas, must be "no". Coherence at the second order does not rule out the possibility that one could be heteronomous with respect to one's volitions. One could have appropriated one's second-order attitudes uncritically from others, or they could have been formed as a result of manipulation or deception.

There are other objections which one could raise against the suggestion that coherence is sufficient for the autonomy of one's desires. For coherence can be attained by revising one's preferences to conform to externally constrained options. Consider, for example, a highly constrained prisoner, whose freedom of action is limited only to pacing back and forth over a five-foot area and making scratches on the wall. If one were to take coherence as sufficient for desire autonomy, then the convict would attain the greatest degree of autonomy possible by forming a second-order volition to act only on the desires he is in fact free to act on, and rejecting all others he might have.[8] This conclusion itself is unsettling. But now suppose that the prisoner is set free, and that he retains precisely the

---

[8] This is a well-known objection to wholly internal accounts of freedom of the will or autonomy. In important respects it is the same worry that the "happy slave" or the satisfied inhabitants of Brave New World raise; it can also be considered as a case of adaptive preference-formation of the type analyzed by Jon Elster in Sour Grapes (Cambridge University Press, 1983), especially Chapter 3. The question concerning the importance of open options will be taken up more thoroughly in Chapter VII.

same set of desires he developed while incarcerated. It is true that he is free from internal conflicts and by all accounts he does what he most wants to do when he paces back and forth over an area of just five feet. His desires are, then, authentic. Yet no one, I presume, would call such behaviour autonomous. The question is, why not? An answer can only be given by considering, not just the internal state of the person's desires, but the context in which those desires are formed, evaluated, maintained or rejected. A principal feature of that context must be the options that the person is free to pursue, and the person's beliefs about what those options are.

The prisoner in our example highlights a condition of autonomy which, to my knowledge, has not been remarked upon, namely, that autonomy should be preserved in the face of increased options. Autonomy must be monotonic; that is, autonomy is a monotonic relation.[9] To say that autonomy is monotonic is just to say that if a desire was autonomous at time t_1, given the options $\{a, b, c\}$, it should still be autonomous at some future time t_2, if the set of options is increased to include $\{a, b, c, d, e\}$, provided that the only difference between t_1 and t_2 is the

---

[9] In claiming that autonomy is monotonic in this sense I mean to draw an analogy between the property of autonomy and that of validity from formal classical logic. In classical logic, the property of validity between a set of premises and a conclusion is monotonic because adding more premises does not destroy validity. That is, validity is preserved if the premise set is expanded. The claim here is that autonomy is preserved if the set of options one faces is expanded. That is, the autonomy of a desire is not destroyed simply by an increase in one's options. This is known as upward monotonicity. I shall argue in Chapter VII that autonomy is also downward monotonic, that is, that it is preserved in the face of restrictions on one's option set as well.

enlarged set of options. If one has decisively committed oneself to a particular desire, then that commitment should be maintainable in the face of increased options *without absurdity*. If we were inclined to say that the prisoner is autonomous in adapting his preferences to his constrained options, then we should also say that he is autonomous if he maintains those preferences once his options are expanded. If we want to conclude that the prisoner is not autonomous once he is freed, then we must also say that he was not autonomous while incarcerated.

In arguing that autonomy is monotonic in this sense I do not mean to deny that increased options will often lead to a revision of desires and commitments. We often contract our desires to conform to external constraints, as a way of reducing frustration, for example, and we often revise our desires in the face of new options becoming open or by coming to see them as salient. The ability to revise one's preferences when one has good and sufficient reasons to do so seems to be a necessary condition of autonomy, indeed, and increased options may often provide good reasons for such revision. I mean to contradict none of these observations, nor do I wish to imply that such revisions must be non-autonomous. The plausibility of these considerations rest upon an unstated assumption, though, namely, that the increased options are explored, experimented with, and lead to a change in desires which is based on the new experiences which one's increased options have made possible. The point I am making here is that just having increased options, without the assumption that

experience of them has resulted in an informed revision of one's desires based on new experiences, should not necessarily make our prior commitments non-autonomous.[10] In the example of the prisoner, his preferences are absurd in the light of increased options; that absurdity functions as a *reductio ad absurdum* against the claim that they were autonomous while he was imprisoned. Coherence models alone cannot accommodate this feature of autonomy, for one can achieve coherence by "scaling down" one's preferences so that they conform to radically limited options, in ways which violate the monotonicity of autonomy.

That coherence can be achieved by revising one's higher-order desires to fit one's lower-order desires, as well as by reforming one's lower-order desires to fit one's reflective desires, is thought to have certain unintuitive implications: that one can achieve greater autonomy by scaling down one's reflective motives implies that a slave can attain autonomy by abandoning any higher-order desires she has that she is unfree to act upon, adopting instead a single higher-order volition to do just what her master wants her to do. The happy slave, whom we visited in the previous chapter, would be autonomous while the discontented slave would not.[11] This worry, first raised by Berlin, has recently been taken up by John Christman, who writes that "if liberty is construed as rational self-mastery, then I am made more free when, instead of removing restraints faced by

---

[10] I expand on this notion of monotonicity in the second section of Chapter VII.

[11] See Berlin 1969, p. 135.

my real wishes, I am manipulated into giving up those wishes."[12] The reference to "manipulation" in the statement of the problem gives one reason to suppose that the process by which one has come to revise one's preferences violates the constraints of procedural independence. But Frankfurt's position is that wholehearted identification with one's desires is sufficient to confer autonomy upon them, and so we cannot appeal to the conditions of procedural independence to block the inference that the slave's revised preferences are autonomous when she gives up those plans and purposes she approved of before the revision, in order to bring her desires in line with her circumstances.

Finally, consider the problem which is posed by someone like the "deferential wife" described by Thomas Hill, Jr..[13] The deferential wife maintains an attitude of subordination and servility with respect to her husband. She does so, not because she believes that it is prudentially wise to do so, nor because she thinks it is instrumental to any goal with which she identifies (because she loves him and believes it will make him happy, for example). Rather, she believes that her own concerns and desires are simply less important than those of her husband. Therefore, she defers to his wishes whenever there is a conflict between what she wants and what he wants.

The deferential wife is hardly a paradigm of an autonomous agent. We need

---

[12] John Christman, "Liberalism and Individual Positive Freedom", Ethics 101 (Jan. 1991), p. 352.

[13] Thomas Hill Jr., "Servility and Self-Respect", The Monist 57 (1973), p. 88.

not suppose, however, that she has any unresolved conflicts among her desires, particularly at the second-order. Her self-conception and her specific desires may be in thorough accord. Her servile desires pass the test of authenticity. Yet she seems, at least to many, to be a non-autonomous person and certainly heteronomous with respect to most of her particular desires and volitions.

These cases, together with our observations concerning ambivalence, allow us to clarify the relationship between coherence and autonomy and authenticity and autonomy. Coherence among one's first-order desires is not necessary for agent autonomy; coherence is more important at the level of one's volitions, but even here some inconsistency can be tolerated without undermining agent autonomy. Coherence at the first-order is not sufficient for agent autonomy, since the latter has been characterized as the capacity to form second-order volitions; the happy slave makes it clear that coherence at the level of volitions is also not sufficient to establish agent autonomy. The fact that one can have heteronomous or servile volitions which are nonetheless coherent illustrates that coherence is not sufficient for desire or act autonomy, either, though here coherence is necessary. It must be concluded, then, that coherence is not sufficient for autonomy in any form.

Identification with a desire, in the absence of conflict, is sufficient for its authenticity, however. Insofar as wholehearted identification with a desire ensures that it is truly one's own, internal to one and partly constitutive of one's self-conception (i.e., that it is authentic as we have been using the term) but does not guarantee that the desire is autonomous, authenticity is not sufficient for

desire autonomy. Thus, whether we take the bi-level account of desires as providing a theory of authenticity or of coherence, we must conclude that it falls short of providing an adequate account of autonomy. Given that its central proponents have presented the bi-level theory as a theory *of autonomy*, however, we shall examine this conclusion more fully in the next chapter.

CHAPTER IV

INTERNALIST VS EXTERNALIST CONCEPTIONS OF AUTONOMY

IV.1 INTRODUCTION

The attainment of personal autonomy, as I have characterized it in Part I, depends upon the development of the cognitive, conative and affective capacities needed for the critical evaluation of one's own reasons for action. The bi-level theory has been shown to have considerable strength in its account of these capacities. Yet there remain deep divisions among philosophers concerning how to characterize the competency which makes autonomy possible, even among those who accept the bi-level theory of desires and its relevance for personal autonomy. I shall mark what I take to be the central division with the names "internalism" and "externalism". Internalist theories of autonomy are committed to the view that autonomy is to be defined just in terms of the subjective attitudes of an agent toward her desires (and, perhaps as well, toward the processes by which they have been formed). Internalists hold, that is, that autonomy is wholly a function of the psychological states of the individual. Thus what is essential to what I call the internalist position is the view that the individual's own attitude is the final arbiter of autonomy. Identification with a desire is taken as sufficient to confer autonomy upon it. This identification may itself be subject to some idealizing conditions (e.g., it must be "wholehearted", or directed toward a desire such that

one's approval of it would survive disclosure of its origins). But on an internalist view, autonomy is not dependent on any objective feature of the agent or her situation. Externalist theories, by contrast, deny that identification with a desire is sufficient to confer autonomy upon it. Some condition or conditions "external" to the agent's attitudes are necessary conditions of autonomy. In an externalist theory, the autonomy of a desire depends, for example, on its rationality or on its alignment with the agent's objective interests. Thus externalists impose some idealizing conditions upon the exercise of critical competence and the act of identification which depend for their justification upon considerations other than the actual desires and other subjective attitudes of the agent.

In this chapter I will explicate the internalist conception of critical competence. The bi-level theory of desires will provide the background to this discussion, of course, but here we shall be concerned with examining an internalist explication of the condition of identification. I shall draw particularly from the works of Frankfurt and Christman in presenting the core internalist commitment. What unites internalist conceptions of autonomy is the insistence that autonomy is ultimately conferred upon a desire by the subjective attitudes of approval of the agent whose desire it is, as well as the agreement that, at least in principle, any desire could be an object of autonomy-conferring approval, regardless of its content.

Externalist criticisms of the internalist understanding of autonomy-conferring identification will then be explored. The externalist position can be seen in the

writings of Kant, as well as those of Hill, Campbell and Babbitt.

I will argue that the internalist conception of critical competency is unacceptable, and so some version of externalism must be correct. I will also argue that the externalist theories which have been offered are themselves problematic. It will be my objective in the next three chapters to develop a more adequate externalist conception of autonomy.

IV.2 IS IDENTIFICATION WITH A DESIRE SUFFICIENT TO CONFER AUTONOMY UPON IT?

I shall examine two approval-based theories of autonomy, which share the core internalist assumption just described as well as a commitment to the bi-level theory of desires, though they differ in other important respects from one another. The positions presented are taken from the writings of Harry Frankfurt and John Christman.

FRANKFURT: Frankfurt's argument, that a desire is autonomous just in case the person whose desire it is wholeheartedly identifies with it after reflecting upon it, represents the core internalist commitment: he explicates the autonomy of a desire just in terms of the attitudes of the agent towards it. Thus he writes, for example, that "a person's approval of a passion that occurs in his history is a

sufficient condition of the passion's being internal to him."[1] In "Identification and Wholeheartedness", Frankfurt explains why it is that identification with a desire makes the desire internal to the person whose desire it is, integral to him in such a way that it has a kind of authority which is denied to other desires with which the agent has not identified. Insofar as identification is "decisive" and "wholehearted", the desire so identified with can be attributed to the person as his own, for such decisiveness makes further evaluation unnecessary. "For a commitment is decisive if and only if it is made without reservation, and making a commitment without reservation means that the person who makes it does so in the belief that no further accurate inquiry would require him to change his mind. It is therefore pointless to pursue the inquiry any further."[2] In this way a decisive commitment "resounds" throughout the potentially unlimited sequence of possible further reflections which the agent could engage in concerning the desire and his approval or disapproval of it, making termination of the reflective sequence non-arbitrary.

> Terminating the sequence at that point -- the point at which there is no conflict or doubt -- is not arbitrary. For the only reason to continue the sequence would be to cope with an actual conflict or with the possibility

---

[1] Frankfurt 1977, p. 64. Let me again remind the reader that Frankfurt is using "internal" and "external" here as a property of desires. A desire which is internal, in Frankfurt's sense, is one which the agent has identified with and it is not only authentic but autonomous as well. His usage should not be confused with the distinction I am developing between internalist and externalist theories of autonomy.

[2] Frankfurt 1987, pp. 168-169.

> that a conflict might occur. Given that the person does not have this reason to continue, it is hardly arbitrary for him to stop.[3]

Thus Frankfurt concludes that identification with a desire, provided that the identification is wholehearted in this sense, is sufficient to confer autonomy upon it. While one might accept that the (reasonable) belief that no further reflection is necessary allows a non-arbitrary termination of the higher-order evaluation of one's desires, Frankfurt's view is still susceptible to very powerful counter-arguments.

Most importantly, the possibility of heteronomy at the second order undermines any claim that autonomy is conferred upon a desire just in virtue of its being endorsed at the second order.[4] Someone could, as a result of fierce conditioning and manipulative socialization, come not only to have, but also to identify with, values and desires which she has been manipulated into having. This shows that identification with a desire is not sufficient to confer autonomy upon it.

Consider, by way of illustration, the example of a contented but subservient housewife in an extremely sexist patriarchical society. Given the dominant values

---

[3] Frankfurt 1987, p. 169.

[4] Campbell raised this objection to the bi-level view very early in Self-Love and Self-Respect, pp. 148, 212-214, as did Irving Thalberg in "Hierarchical Analyses of Unfree Action", Canadian Journal of Philosophy Vol. VIII, No. 2 (1978). See also Watson 1975 and Wolf 1988. This challenge is taken up in the debate between Friedman 1986 and Christman 1987. Christman's position is further developed in "Autonomy and Personal History, Canadian Journal of Philosophy Vol. XXI, No. 1 (1991); hereafter Christman 1991b.

which we can imagine her to have been socialized to accept, we might suppose that she would develop certain pre-reflective first-order desires: the desire to get married, the desire to make her husband happy, the desire to have children. She would also be expected not to develop certain other desires: the desire to pursue an advanced education or a career. Given that the society in which such a person is raised leaves some options open to her -- that marriage is not arranged by families during early childhood, that education and some career options are available to women and the like -- it is possible that she could develop the capacities needed for reflection. Yet it is also plausible, at least, to suppose that she would endorse her subservient desires at the higher order, were she to reflect upon them. Furthermore, there is no reason to suppose that her attitudes of approval toward "the roles and desires appropriate for women", and her disapproval of more independent projects or desires, would not be decisive, in Frankfurt's sense. She might experience no conflict in her desires at any order and she might reasonably believe that there is no further need for reflection or doubt concerning what she really wants to do. Her real self and her ideal self may be in complete accord. If she feels any desires in the future which vary with that conception of "a good woman" or which conflict with the desires of her husband, she will disapprove of them and reject them upon reflection. If driving a car is inconsistent with this self-image or her husband's wishes, for example, then should this desire manifest itself she will simply reject it. Yet such a woman seems to many people to be heteronomous, having merely adopted her desires

and values from her society in such a way that they cannot be autonomous. Given that her desires are acquired within a system of systemic oppression and that the contents of her desires are self-effacing and incompatible with genuine self-respect,[5] she seems hardly to be a paradigm of an autonomous agent. Yet, on Frankfurt's view, we have no grounds for challenging the claim that her subservient desires are fully autonomous. This type of counter-example is controversial, however, and even feminists are divided over the question of whether or not a woman could autonomously choose the role of a traditional wife under patriarchy,[6] and so I shall develop another.

Consider now the case of a person who has been subjected to the kind of conditioning techniques that children in Aldous Huxley's Brave New World are subjected to, particularly the relentless sleep-conditioning techniques through which they are indoctrinated to acquire socially approved-of class, work, leisure and sexual values. The inhabitants of Brave New World feel no conflict among their desires; if they reflect upon them at all (and some do), they can be expected to decisively identify with them. Again, they seem to pass Frankfurt's test and so we should conclude that their desires are autonomous. Yet the contented inhabitants of Brave New World seem paradigms of heteronomous, manipulated

---

[5] I shall assume that her desires are incompatible with genuine self-respect here without argument.

[6] For a discussion of this debate, see Meyers 1987 and "The Socialized Individual and Individual Autonomy", in Woman and Moral Theory eds. Feder Kittay and Diana Meyers (Rowman & Littlefield Publishers, 1987); the latter will be cited hereafter as Meyers 1987b.

individuals.

Suppose, finally, that there is a particular politician in a small town, who is sincere and serves the interests of her constituents admirably and tirelessly for the most part, but who used bribery to attain her political office. If this past breach were to be made public, she would be forced to resign her political post and so would be unable to bring about a number of worthwhile policies which she can pursue in her political capacity. Suppose, furthermore, that her past transgression should become known to another, who then uses it to blackmail her. The other threatens her with exposure unless she pays the former a certain sum of money. Upon serious reflection the politician may decide to pay the blackmail money. If the threat is credible and she reasonably believes that paying will ensure her blackmailer's silence, then she may decide that it is best on balance to pay the money. If she then experiences renegade desires, to resist her coercer, she may want to resist them, and to cultivate a volition to act on the desire to pay the money. Alternatively, fearing that she will be overcome by such renegade desires, she may decide to reject her desires to resist as good reasons for action, forming a volition not to act on them. Insofar as the formation of these volitions is undertaken because they serve what she believes is best, all things considered, to do, and she has no reason to question that judgment further, she may identify decisively with her desire to pay the money. On Frankfurt's view, her decision to comply with the threat must be seen as autonomous. But she is acting on the basis of a coercive threat, and many people believe that desires which are caused

by threats cannot be autonomous; indeed, there is an intuitive sense in which all actions which are motivated by a coercive threat are paradigmatic cases of heteronomy, or direction by another, rather than self-direction.

These examples suggest that heteronomy at the second order is a real possibility, and that neither coherence at the level of one's volitions nor wholehearted identification is sufficient to ensure that the processes of desire-evaluation and identification have not been interfered with in ways that undermine the autonomy of the desires identified with. Critical competency cannot, then, involve just the capacities necessary for decisively identifying with a desire, for exercising those capacities is not sufficient to confer autonomy upon the desires thus identified with.

The basic problem posed by the possibility of heteronomy at the level of one's second-order volitions, then, is this: autonomy is supposed, on the bi-level theory of desires, to be conferred by the process of identifying with one's motives at the second order. But unless that process of identification is itself autonomous, it is unclear why identification has the autonomy-conferring status it must have. If the process by which identification confers autonomy on lower-order desires is itself autonomous, on the other hand, then we need an account of the autonomy of this process. If it is conferred by reflection upon and approval of the act of identification itself (or the second-order desire that results), then we have a

serious regress problem.[7] In this form the regress problem is a substantial challenge for any internalist bi-level theory which claims that second-order identification with a desire is sufficient to confer autonomy on it. For if one's first-order desires are truly one's own because one has identified with them at the second level of desire-formation, then we can raise the question whether or not one's second-order desires are truly one's own. We answer this question concerning first-order desires by determining whether or not they have been endorsed at the second order. To answer the same question concerning our second-order desires, then, perhaps we have to consult our third-order desires, etc.. That is, one might have to posit higher and higher orders of desire-formation in order to ensure that one's highest-order desire is truly that which one approves of.

The foundation for this regress challenge is the internalist premise that identification must be specified just in terms of attitudes of the reflective agent. Yet Frankfurt recognizes that any account which claims that identification is sufficient to confer autonomy upon the desire identified with, if identification is cashed out just in terms of attitudes of approval or resignation of the agent, is subject to the charge of generating an infinite regress of reflection. Thus Frankfurt writes,

> [T]here is a quite basic error in thinking that the concepts of internality and externality are to be explicated <u>simply</u> in

[7] See Christman 1987, pp. 283-284, for an interesting discussion of this problem and its various solutions.

> terms of a person's attitudes. It is fundamentally misguided to suggest that a passion's externality is entailed by the person's disapproval of it, or that its internality is entailed by his approval. The trouble with this approach to the problem of understanding internality and externality is that it fails to take into account the fact that attitudes toward passions are as susceptible to externality as are the passions themselves... The fact that a person has a certain attitude toward a passion can be construed as determining either the internality or the externality of the passion, surely, only if the attitude in question is itself genuinely attributable to him. An attitude in virtue of which a passion is internal, or in virtue of which a passion is external, cannot be merely an attitude that a person finds within himself; it must be one with which he is to be identified. But given that the question of attribution arises not only with regard to a person's passions, but also with regard to his attitudes toward his passions, an infinite regress will be generated by any attempt to account for internality or externality in terms of attitudes. For the attitude that is invoked to account for the status of the passion will have to be an internal one; its internality will have to be accounted for by invoking a higher-order attitude -- that is, an attitude toward an attitude; and so on. [8]

I have quoted Frankfurt at length here, partly because his is a clear statement of the problem and partly because it is remarkable that he develops the challenge so fully while admitting that he has no answer to it. Thus he recognizes that endorsement of a desire at the second order is not sufficient to ensure its internality, nor is disapproval sufficient to guarantee externality.

One way of avoiding this problem, of course, would be to abandon the

---

[8] Frankfurt 1977, pp. 65-66. Here Frankfurt is again contrasting desires that are "internal" to a person and those which are "external" to her in order to mark the difference between authentic and unauthentic desires. So used, internality or externality is a property of desires. This distinction ought not to be confused with the distinction I am drawing in this chapter between internalist and externalist theories of autonomy.

requirement that one's second-order desires must themselves be autonomous. This is clearly not a viable option, however, for it would expose the theory to an *ab initio* problem: second-order approval would confer autonomy on first-order desires "from nowhere". There would, then, be no basis for the claim that second-order endorsement of a desire confers autonomy upon it. Furthermore, taking this option would be equivalent to saying that identification with a desire is sufficient for its autonomy, and so would open the theory up to the counter-examples developed just above.

Luckily, we are not forced into the *ab initio* problem to avoid the regress challenge. For the regress challenge gets off the ground only if we suppose that second-order desires (or the acts of reflection and identification which lead to their formation) must be autonomous in the same way as first-order desires are autonomous within the bi-level theory, that is, by being identified with at a higher order. To avoid the regress and *ab initio* problems, we must conclude that second-order desires are autonomous in a different way from first-order desires within bi-level theories. Insofar as the theory provided by Frankfurt fails to provide an account of autonomy at the second order, it is seriously incomplete. But we already realized this: it follows directly from the conclusion that identification with a desire is not sufficient to guarantee that it is autonomous.

Christman thinks that the problems enumerated here are unavoidable because he denies that identification can be specified independently of the attitudes of the

reflective agent.[9] He claims that all self-appraisal models of autonomy rests on the following premise: "that the only account of the authenticity of the acts of appraisal that comprise autonomy must refer to other preferences of the agent."[10]

CHRISTMAN: Christman has developed a more sophisticated internalist conception of autonomy, which he believes is immune from the criticisms to which Frankfurt's analysis is open. Christman's proposal remains faithful to the self-appraisal model of autonomy, however, and so is internalist.

Christman's basic position is that, because it is not enough that one approve of one's first-order desires to ensure that they are autonomous, one's approval must not be caused by "illegitimate external influences": second-order identification with (or approval of) a desire is itself autonomous just in case the identification was not caused by illegitimate external influences.[11] Illegitimate external influences are external factors which interfere with autonomous higher-order desire-formation: they must be specific to the higher-order desire and explanatorily adequate to account for its formation. Furthermore, the influence of such factors must be such that, were the person to attend to them, he would

---

[9] Christman 1991b, p. 8

[10] Christman 1991b, p. 18.

[11] Christman believes that this characterization of autonomous desires allows him to avoid the regress problem, for he claims to be citing both necessary and sufficient conditions for desire autonomy.

resist their influence. What the agent must approve of, then, is not only a given desire, but also the process by which the desire was formed and endorsed at the second order.

> The acts of critical reflection and identification with LOD's [lower order desires] must not themselves be caused by Illegitimate External Influences. Such factors are ones which arise essentially from outside of the person's normal channels of cognitive processing. And, were the agent to be made aware of their presence and influence, she would be moved to revise her desire set.[12]

This suggests that there are two means of identifying an illegitimate external influence: it arises outside of one's normal cognitive processes and one subject to it would resist its influence if one were aware of it. Christman says that these are independent tests, and his discussion implies that either is sufficient to make an influence illegitimate.[13]

Christman has offered a number of different formulations of the kind of test he envisages, which one's reflective desires must meet if they are to count as genuinely autonomous. He emphasizes the first of these tests in his 1987 paper, arguing that illegitimate external influences are identifiable insofar as they are causal influences "originating from outside of her normal cognitive processes".[14] While we might complain that this condition is too vague to be of much service, I shall suppose that it is at least adequate to deal with preference changes which

---

[12] Christman 1987, p. 291.

[13] Christman 1987, p. 290.

[14] Christman 1987, p. 290.

result from such processes as drugs or hypnosis. This itself may be problematic, however, for it is widely recognized that one can autonomously choose to undergo hypnotic or drug therapy in order to induce a preference change which is desired; a frequently cited example is of a person who undergoes hypnosis in order to quit smoking. Presumably such means are outside of one's normal cognitive processes, though they do not defeat one's autonomy. Even if we ignore this problem, though, I suggest that Christman's own example poses another challenge to our understanding of this test. His example is as follows.

> An elderly person who is terminally ill and in great pain refuses a fairly minor operation that would prolong her life for another year.... Following Dworkin, we determine [whether or not her decision was autonomous] by asking whether there is identification of the right sort with the desire to forego the operation (the LOD) -- i.e., whether the desire is authentic. On the addition to that account which I am here suggesting, we must also ask if any external factors are essentially the cause of that identification, factors not part of what the woman identifies as herself. Imagine that her son has made various remarks concerning the expense of the hospital stay. The woman may, then, be caused to approve of her own desire to die as a direct result of this external pressure. If this were so, we would not regard her decision as fully autonomous.[15]

This case strikes me as extremely puzzling in the context of Christman's discussion. On the face of it, at least, the son's remarks concerning the expense of hospital care do not seem to be the kind of influence which might be thought of as influencing his mother "from outside of her normal cognitive processes". Indeed, for such "information" to have any influence on her at all, one might

---

[15] Christman 1987, p. 290.

think, it would have to be brought within her cognitive processes. Moreover, on plausible theories of practical deliberation, the woman would also have to have some desire or goal which this fact is relevant to: the desire not to be a burden to her son unnecessarily, or the goal of leaving him financially well-off, might be likely candidates here. Perhaps what Christman means, when he says that the external factors are "not part of what the woman identifies as herself", is that these desires are not ones with which she identifies. But if so, then how does the knowledge that the operation will be costly influence her decision at all? If she simply did not care about her son's financial well-being, or had rejected a desire for his financial security as a reason for action, then it is difficult to understand how his repeatedly mentioning the cost could influence her at all. If Christman just means that she has not reflected upon these desires, though she has them, then the story seems to support the claim that she would identify with them if she thought about them as the objects of evaluation; this, at the very least, is a possibility. Christman's claim that "A person is rendered less autonomous when specific and identifiable factors, originating from outside of her normal cognitive processes, cause a change in the preference structure of the person"[16] gains no support from such a case.[17]

---

[16] Christman 1987, p. 290.

[17] Though I think that Christman has failed to show what, if anything, is autonomy-destroying about the woman's worry concerning the cost of her medical care, I do not mean to imply that those who discuss autonomy and consent in the context of medical ethics are misguided in worrying, as they do, about such subtle forms of influence which can be brought to bear on patients by

Perhaps it is just nitpicking to argue against what is supposed to be a general claim about autonomy by showing that the example Christman uses is inadequate to support it. Let us, then, examine his claim directly. Is it necessary to subvert the normal cognitive processes of an agent to defeat her autonomy? No. One can do that by inducing false instrumental desires (by withholding relevant information from the agent, say) without by-passing the person's normal cognitive processes. As Stanley Benn has pointed out, "There is no point in deception and censorship unless the subject can be expected to form his beliefs on evidence, and to act on them; otherwise providing false evidence would not be a way of controlling action."[18] Likewise, the son's actions in Christman's example can influence his mother's behaviour or desires only on the assumption that her cognitive and affective processes are functioning normally.

Is it sufficient to defeat autonomy that one employ influences which operate outside of the agent's normal cognitive processes? No, again. Examples of someone autonomously choosing to undergo hypnosis or psychosurgery are sufficient to show this. The first of Christman's tests for illegitimacy of an external influence is not relevant to questions of autonomy, then.

We must turn to the other test for the illegitimacy of external factors which

---

their families and health care professionals. My point, here, is just that the patient must have some commitments to be exploited for such forms of pressure to influence her. The fact that they can be exploited tells us nothing about whether those commitments are, themselves, autonomous.

[18] Stanley I. Benn, "Freedom, Autonomy and the Concept of a Person", Proceedings of the Aristotelian Society (Jan. 1976), p. 112.

Christman suggests, which concerns the attitude the agent would take toward those factors were she to attend to their influence in the formation or revision of her motivations. This is the condition that does the work in Christman's final theory, and it is here that his internalist commitments are at the fore. For the autonomy of one's desires will depend just on the attitudes one has towards the processes by which those desires were formed or endorsed.

> [A]ny factor affecting some agent's acts of reflection and identification is "illegitimate" if the agent would be moved to revise the desire so affected, were she to be aware of that factor's presence and influence. That is to say, if an agent comes to know that a certain factor (hypnosis, for example) played a crucial role in the formation of, and identification with, a certain preference, and she revises her approval of that preference as a result, then the factor is considered illegitimate. Hence, that desire would be considered non-autonomous.[19]

This seems a more promising strategy, for it allows that one may choose to submit to methods of influence (such as hypnosis) which operate outside of normal cognitive channels without forfeiting one's autonomy. Presumably, if one autonomously chose to submit to such influences, then coming to know that one's present desire-structure was caused by those influences would not lead one to revise that structure.

Other forms of influence, such as the manipulation of information, could well be used in such a way as to cause a person to identify with a desire which she would not have identified with but for the manipulation. Should such an

---

[19] Christman 1987, pp. 290-291.

influence become known to her, she might well be expected to revise her preferences in light of the new information.

Christman's formulation of this test is not perfectly clear, however. As it is stated, it seems that the propensity to revise one's preference in the light of new information concerning the influences causally responsible for one's coming to approve of it is sufficient to declare the desire non-autonomous. Is the presumption that, if such disclosure is made and the person does not revise her desire structure, then the desire is autonomous? This is certainly his position in "Autonomy and Personal History", where he writes that "If the act of appraisal of the processes by which a desire developed in an agent is carried out with sufficient self-awareness and minimal rationality then that act of appraisal (and non-resistance) is sufficient for the autonomy of the desire."[20] This seems problematic, in the face of cases like that of the subservient housewife. She may have originally adopted her subservient desires and identified with them because of external influences (manipulation of information and restrictions of options for women, having been educated under patriarchy to accept certain false beliefs about the proper role and values for women, which beliefs might then have been reinforced by the mass media and religious authorities of her culture, etc.); but even if she comes to have full information about the role of such influences in causing her to approve of her subservient desires, she may persist in them anyway.

---

[20] Christman 1991b, pp. 18-19.

The problem posed by persistent identifications which survive the disclosure of the processes by which they were adopted, as highlighted by examples of subservience, is anticipated by Christman. Consider his own version of this example:

> Imagine ... a woman who is raised in a culture which fiercely inculcates in her the idea that women should never aspire to be anything but subservient and humble domestic companions to their husbands, no matter how unhappy this makes them or how abusive their husbands are. Imagine further that this person is suddenly placed in a new culture where opportunities abound for women to pursue independent activities. She nevertheless shuns these opportunities and remains married to an oppressive husband from the old culture. The only "restraint" she faces (to pursuing the opportunities for an independent life-style) are her desires themselves (which remain the sort she was taught to have). She simply does not wish to act in any other way, turning a deaf ear to the reasons people give her to consider a less subservient posture.[21]

Clearly the processes by which such a woman came to have the desires and values she has were oppressive and, we can suppose, such that they "did not allow her to reflect on her emerging values in light of reasonable alternatives". Yet, as Christman's description of the case shows, it does not seem impossible to imagine that those desires and values, as well as the behaviour they inspire, could persist once she comes to understand how she came to form them and when she is exposed to more attractive open options. The problem is that what one might accept will vary depending on the commitments one has made. If one has accepted and internalized a particular value, then even disclosure of the processes

---

[21] Christman 1991, pp. 344-345.

leading to its adoption may not be sufficient to dislodge it after the fact.[22] Thus, Christman concludes, if we are inclined to say that, despite her identifications with her desires in the face of knowledge of their origins, th subservient housewife is not autonomous with respect to them, then we will have to go beyond this second, internalist, criterion of "illegitimacy" for methods of influence.

Christman thus recognizes that some further constraint is needed on the processes of desire-formation in order to bar such possibilities as those raised just above. He argues that the problem here is that the subservient woman's desires and values have been oppressively imposed upon her, and concludes that the processes by which one comes to form one's desires must be such that one had (or could have had) some control over.

> Preference changes cannot be the result of oppressive conditions or blind, unreflective conformity to limited choices. Self-mastery means more than having a certain attitude toward one's desires at a time. It means in addition that one's values were formed in a manner or by a process that one had (or could have had) something to say about.[23]

This condition, which I shall refer to as the condition of control, is problematic, however. Christman's position is that if one's desires were formed by processes over which one had no control or no say then one cannot be autonomous with

---

[22] We cannot appeal at this point to the counterfactual component in Christman's proposal, for that would make his test counterfactual in principle: a desire is autonomous only if one would have accepted its formation, knowing the processes by which it would be formed, before it is adopted. This would make the test inapplicable to too many cases for it to be of use.

[23] Christman 1991, p. 346.

respect to them. This is troubling, for presumably most of our early education and socialization proceeds via processes over which those subject to them have no say.[24] Another example may bring this out. Imagine a woman raised in a cultural setting which is truly committed to gender-equality, in which there are truly open options for both men and women, etc.. Suppose, furthermore, that her parents conscientiously instilled in her a healthy sense of self-respect and appreciation of her talents and the open options which she faces. This woman might be expected to develop values and desires which reflect the influences to which she was exposed: she might desire only mutually supportive and empowering relationships, to pursue a particular career, etc.. Yet, insofar as her desires and values have been formed because of her parental and societal influences, one might think that they were formed by processes over which she had (or could have had) no control, "no say about". Must we conclude, then, that she is not autonomous with respect to the desires which have developed as a result of this process?

Here, recent feminist insights are relevant. Sarah Hoagland, in Lesbian Ethics, advocates abandoning the traditional conception of autonomy and adopting

---

[24] This is, doubtless, too strong. As David Braybrooke has reminded me, even very young children have "some say" in the processes of their socialization: they do resist various lessons, sometimes successfully. Nonetheless, the sort of control over the processes of one's education and socialization that Christman here claims is necessary for autonomy -- that the agent was in a position to reflect on and resist the factors which produce a change in her preference set -- is such that no one has much of it during their childhood. Christman 1991, p. 346.

instead a model of the self which is "autokoenonous". Hoagland's description of this alternative ideal will be unsatisfyingly vague for analytic philosophers: "a self which is both separate and related, a self which is neither autonomous nor dissolved: a self in community who is one among many."[25] I think that what she has in mind here can be made clearer, and points to a deep problem with Christman's view. For she wants to reject the view that individuals acquire autonomy only when they acquire not only control of themselves, but control over their external circumstances as well. Once we come to appreciate that our position within communities makes these conditions of control impossible to achieve, we shall have to abandon such an ideal of autonomy. The feature which differentiates autokoenony from autonomy is that the former depends upon the ability to make choices within oppressive situations, the ability to maintain one's integrity and avoid being demoralized by oppression and to connect with others in ways which are empowering. This requires acting and choosing and valuing in whatever situations we find ourselves.[26] This is the mark of moral agency. "My suggestion," writes Hoagland, "is that moral agency involves enacting choice in limited situations, avoiding demoralization, and working within the boundaries rather than trying to rise above them."[27]

The reality of coercion and oppression of women creates serious problems for

---

[25] Sarah Hoagland, Lesbian Ethics: Toward New Value (Institute of Lesbian Studies, 1988), p. 12.

[26] Hoagland 1988, pp. 144-145.

[27] Hoagland 1988, p. 198.

those who want to increase the scope of female autonomy, particularly when the coercion is so systematic that its victims internalize the oppressive values of their coercers. But this is not just a problem for women or other groups who are systematically oppressed or exploited: everyone finds himself or herself enmeshed in social practices over which they had and have no control (or, very little control). Hoagland notes that under traditional conceptions of autonomy (and I take it that Christman's falls into this category), one acts autonomously only if one acts as one wants in situations in which one has a genuinely open choice, both concerning one's participation in the various practices which are operative, and with respect to whether one will internalize the values of these practices or not. It is for this reason that Hoagland wants to reject such conceptions, for they rest on an unrealistically simple representation of the actual choice situations in which moral agents must act. To be autokoenonous is to maintain one's moral agency in less than ideal choice situations, including situations of oppression, to continue to act and create value by one's choices:

> My thesis is that moral agency simply is the ability to choose in limited situations, to pursue one possibility rather than another, to thereby create value through what we choose, and to conceive of ourselves as ones who are able to and do make choices -- and thus as ones who are able to make a difference for ourselves and each other in this living. Moral agents are autokoenonous beings.[28]

Drawing on Hoagland's insights here, I do not want to follow Christman in denying autonomy to individuals just because their preferences and plans have

---

[28] Hoagland 1988, p. 231.

been formed under conditions of oppression or limited choices. To do so would be to conceive of them as "victims of circumstance" and to deny them the status of moral agents, a judgment which would most likely lead to their continued subordination. *No one* has the control that Christman is seeking here, and so we should reject this specification of autonomous preference-formation. If having such control is the only way to bar objections to the internalist, historical test for distinguishing autonomy-conferring acts of identification from those that do not confer autonomy, then those objections stand against Christman's internalist theory of autonomy.

What unites the internalist theories discussed here is that autonomy is conferred upon desires through their being approved of by the reflective agent whose desires they are. As we have seen, there are a range of positions within the internalist camp, concerning when or under what conditions an agent's approval confers autonomy upon a given desire or end. At his most extreme, Frankfurt insists that present approval is sufficient, with no constraints placed upon the context or process leading to that approval. In his more careful moments, he imposes the idealizing condition that the approval be wholehearted. Christman extends the idealizing conditions even further, so that approval confers autonomy upon one's desires provided that the person was minimally rational and had full information concerning the processes by which the desire was acquired.

One thing that is of special interest in these theories is their shared conviction that desires which are approved of, under the specified conditions, represent the real, autonomous interests of the agent. For Frankfurt, a person's real interests are those which she would decisively endorse. For Christman, a person's real interests are those she would identify with if she had full information concerning the causal origins of her desires. Autonomous action will be action which aims to realize one's real interests, so conceived.

The problem with all such theories about what constitutes the real interests of agents is that they take as their basis the actual desires and commitments of individuals. The actual desires of agents provide the point of view from which the idealizing conditions are specified. That is why we said of the subservient housewife that she could be autonomous on the internalist theory, because (given her actual frame of reference and values), even knowing how she came to have her subservient desires may not motivate her to reject them. I shall call the idealizing conditions that the internalists impose "subjective idealizing conditions", for they are specified by reference only to the subjective attitudes of the agent (wholehearted identification or approval in light of full knowledge concerning the etiology of one's desires). One need not, now, value one's own autonomy, and so (even under the subjective idealizing conditions) one could identify with, and so confer authenticity upon, subservient or slavish desires. This is because all internalist conceptions of autonomy are "content-neutral" -- a feature of their theories that Frankfurt and Christman regard as speaking in their favour.

Christman's statement of the content-neutrality of the internalist position is striking: "For any desire, no matter how evil, self-sacrificing, or slavish it might be, we can imagine cases where, given the conditions faced, an agent would have good reason to have such a desire."[29] Provided that one has identified with one's evil or slavish desires, under idealizing conditions which are themselves specified only by reference to one's present frame of reference and values, one's desires are autonomous regardless of their content. Imposing only subjective idealizing conditions upon the autonomy-conferring process of approval forces the internalist to recognize "autonomous slaves" as a genuine possibility. This conclusion ought to function as a *reductio ad absurdum* of internalist theories of autonomy, which attempt to define autonomy just in terms of the subjective attitudes of agents.

IV.3 SUBSTANTIVE EXTERNALIST THEORIES OF AUTONOMY

It is precisely on the question of how to characterize the real interests of autonomous agents that externalists depart from internalists. For externalists hold that individuals may have objective interests, which they ought rationally to endorse and pursue, even if they do not, in fact, desire to do so, and even if they would not desire to do so under the idealized conditions of choice imposed by

---

[29] Christman 1991, p. 359.

the internalists.[30]

Immanuel Kant is a paradigmatic representative of the externalist position. For Kant, autonomy is a property of all rational wills. Insofar as a person is rational, he will acknowledge certain obligations as unconditional, universal requirements of reason; insofar as he is autonomous, he will accept these obligations as binding upon himself. What are these obligations? The categorical imperative, interpreted as the requirement to respect ourselves as well as others as ends in themselves, is the most central in generating these obligations. Since these obligations are prescribed by reason, they constitute objective interests for all autonomous agents, independently of their particular desires or attitudes. As Thomas Hill Jr. explains, Kantian autonomy "includes the idea that rational agents have reasons not based on their desires, that practical rationality is not exhausted by hypothetical imperatives."[31] As such, there are ends which are rational for all autonomous agents, regardless of whether they desire or value those ends.

A very considerable literature has developed around Kant's theory of autonomy and its relation to his theory of rationality, which it is beyond our present purpose to pursue. Rather than pursue Kantian exegesis further, then, I shall turn to some contemporary externalist theories of autonomy, particularly

---

[30] On such a general characterization, a great variety of externalist positions on interests exist: classical hedonism, theories of needs, etc.. The details of the externalist theories I am interested in will be filled out in what follows.

[31] Thomas Hill Jr., "The Kantian Conception of Autonomy", in The Inner Citadel ed. John Christman, p. 103. See also the discussion of Kantian autonomy by Richard Lindley, Autonomy (Humanities International Press, Inc., 1986), Chapter 2.

those developed by authors who take themselves to be criticizing the version of internalism represented by proponents of the bi-level theory. The externalist position is characterized by a denial that identification with a desire is sufficient to confer autonomy upon it, even under idealizing conditions, if those conditions are determined by the agent's present frame of reference and values. Only if the agent's desires conform to some conditions which are external to her subjective attitudes can identification with them ensure their autonomy. (Indeed, many externalists reject the necessity of identification as well, claiming that a desire is autonomous just in case it meets the external requirements for autonomous desires.[32]) Expressed as a thesis concerning the interests of autonomous agents, externalists hold that an account of objective interests can be given independently of the actual or subjectively defined idealized desires of any particular agent.

Richmond Campbell's theory of autonomy meets this requirement, for he claims that one's real or objective interests (or self-identifying desires) are those one would have and be able to sustain in a coherent way if one were fully aware of and attentive to the relevant truths regarding them.[33] One's actual desires, or those one would have under the idealizing conditions imposed by the internalists, may fail to be autonomous. Nonetheless, Campbell argues that having and pursuing autonomous desires is in the objective interests of all agents. For he

---

[32] Cf. Friedman 1986; Paul Benson, "Autonomy and Oppressive Socialization" Journal of Social Theory and Practice Vol. 17, No. 3 (1991).

[33] Campbell 1979, p. 207.

argues that the pursuit of such autonomous desires is an integral component of human flourishing, and all persons have an interest in attaining a state of human flourishing. This is an interest which people share, moreover, independently of whether they desire the conditions which constitute it or not, that is, whether or not their desires are autonomous.

Likewise, anyone who argues that the development of autonomy is in the objective interests of individuals, whether or not they do or would value it under subjectively defined ideal circumstances, is offering an externalist account of autonomy[34]. For the autonomy or non-autonomy of some desires will be determined independently of the agent's attitudes. Susan Babbitt offers such a position, in response to what she takes to be weaknesses in the internalists' construal of real interests (she refers to what I call the internalist position as the "liberal view"). Babbitt argues that people have objective interests in developing autonomy and self-respect, which are independent of their desires for these conditions. The internalist conception of real interests, even given very strong subjective conditions for idealized choice, cannot acknowledge such objective interests, however, and so is inadequate as a theory of autonomous interests.

Babbitt uses Thomas Hill Jr.'s example of the deferential wife to make her point against the internalists. The deferential wife, as we have seen, identifies

---

[34] This is the position adopted sometimes by John Rawls. Cf. Rawls, "Kantian Constructivism in Moral Theory", Journal of Philosophy 87 (1980), pp. 525-526. But compare, Rawls, A Theory of Justice (Harvard University Press, 1971), pp. 248, 417.

strongly with her subservient desires, and her self-conception is importantly constituted by the deferential position she takes toward her husband. Babbitt imagines the case to be such that -- given her servile self-conception and that her view of herself and the conditions in which she could attain a sense of flourishing, self-respect and autonomy have been "deformed" by oppressive social relations -- she may have no reason, even under subjective ideal-choice conditions, to abandon her deferential posture. Even knowing that most people value greater power, self-respect and autonomy than she has in her relationship with her husband, and even being able to imaginatively entertain alternatives in which she would enjoy these goods to a greater degree, this knowledge simply may not "touch her", given her previous commitments. What she may need, Babbitt suggests, is new experiences. While the propositional knowledge just mentioned may leave her cold, having new experiences could transform her self-conception in such a way as to provide a different background for interpreting the significance of alternate possibilities. She may need to undergo what Babbitt calls a "transformational experience", then, in order to appreciate the real interests which she has in developing autonomy and self-respect.

> In a sense, Hill's Deferential Wife may in fact be right in thinking she is not being personally deprived by acting out her deferential relationship. Given the dependence of her identity on her social situation, she may really have as one of her personal characteristics the feature of being inferior to her husband. If she were better informed she would know that it is in human beings' interests, generally, to pursue a full sense of autonomy. But she may well base her actions on assumptions about her worth and prospects that make this general information

> inapplicable to her situation. However, it is likely that if the Deferential Wife were to act in certain ways, or even were compelled to act in certain ways by circumstances or forceful persuasion, she would *acquire* desires and interests that would change her position and provide her with a different interpretive background. If she were to acquire greater power or self-respect, she would in fact become such that the actual denial of power and control to her is a *personal* deprivation. The personally transformative experiences she undergoes could therefore provide her with a more adequate interpretive background for making choices about her life.[35]

Thus Babbitt rejects the liberal (internalist) theory of autonomy precisely because it takes the individual's given preferences as the foundation of her autonomous interests:

> The problem with the liberal view as an account of autonomy is precisely that it rests on the preservation of the initial individual's perspective; liberal accounts define rational interests in terms of what the *individual* would choose under suitably idealized conditions. The problem for this view is that in cases of ideological oppression and false consciousness, individual autonomy appears to depend importantly upon the disruption of the individual's initial perspective and the bringing about of more adequate self-understanding as a result of social and political action.[36]

Within the theories put forward by Campbell and Babbitt, then, there are some ends which no autonomous agent could choose. For this reason they are not "content-neutral".[37]

---

[35] Susan Babbitt, "Feminism and Rational Interests", unpublished CPA manuscript (1992), p.14.

[36] Babbitt 1992, p. 28.

[37] Paul Benson and Thomas Hill Jr. should also be included in the externalist camp, for they also offer non-content-neutral theories of autonomy.

There are serious problems with this externalist conception of autonomy, however. Three will be presented here. First, the externalist position makes autonomy depend upon an agent's having proper respect for her self-interest -- taken in an external (and hence objective) sense.[38] In arguing that the deferential wife is not autonomous, for example, Babbitt assumes that being subservient is contrary to her objective self-interest. Let us grant that it is true, in virtue of deep psychological considerations which pertain to human nature and the universal interests of human beings, that the wife who desires to be subservient thereby evidences a lack of respect for her real interests. Nothing follows directly from this concerning her autonomy, unless we adopt the premise that *no actions (or desires) which are contrary to one's self-interest are autonomous*. This premise seems implausible. Many individuals (e.g., fire-fighters) autonomously risk their own safety for others, yet it is surely in their self-interest not to intentionally put their physical safety at risk. Parents frequently accept harms to themselves, and so act contrary to their self-interest, for the benefit of their children, without any indication that they are other than autonomous in doing so. Some individuals autonomously inflict harms on themselves for short-term excitement or pleasure. Etc. It would seem, then, that the externalist "solution" to the problem of the subservient housewife depends upon a premise which is itself implausible. Respect for one's objective self-interest does not seem necessary for autonomy.

---

[38] I am grateful to Nathan Brett for helping me to clarify this objection to the externalist position.

Moreover, there are many cases of subservience that do not imply a loss of autonomy or sacrificed self-interest. The soldier who identifies with the goal of winning the war, and recognizes that it is necessary that he subordinate his own wishes to those of his superiors in order to achieve that end, adopts a position of subordination but does not thereby forfeit his autonomy. Likewise, the priest who dedicates his life to serving God must subordinate himself to God's will and those who occupy positions of authority in the church hierarchy, yet the decision to become and remain a priest could nonetheless be autonomous. Here autonomy comports with direction by another, because the reflective endorsement of the agent has not been excluded. This supports the conclusion, drawn by Haworth and Dworkin, that "substantive independence" is not necessary for autonomy.

Dworkin argues against incorporating any substantive constraint in a theory of autonomy:

> [T]here is a tension between autonomy as a purely formal notion (where what one decides for oneself can have any particular content), and autonomy as a substantive notion (where only certain decisions count as retaining autonomy whereas others count as forfeiting it). So the person who decides to do what his community, or guru, or comrades tell him to do cannot on the latter view count as autonomous. Autonomy then seems in conflict with emotional ties to others, with commitments to causes, with authority, tradition, expertise, leadership, and so forth.[39]

In arguing against the necessity of substantive independence for autonomy, Dworkin allows that one can subordinate oneself to a goal, cause or individual

---

[39] Dworkin 1988, p. 12.

without necessarily forfeiting one's autonomy. So long as the decision to do what one's commanding officer or church leader tells one to do is a decision made as a result of critical reflection which has not been subverted by violations of one's procedural independence and one is critically competent, then such a decision does not undermine one's autonomy.[40]

The case of the subservient wife succeeds in being a counter-example to the internalist theory of autonomy not because it is a case of subservience alone, but because we are privy to features of the situation which cast doubt upon the reflective capacities of the woman. (Cf. Christman's and Babbitt's discussion of the case in this chapter; we shall return to this example in the next two chapters as well.)

Furthermore, it is obvious that one could act in accordance with one's objective self-interests (at least, with many of them) without being autonomous. It might be just a happy accident that what one desires is actually what will serve one's objective interests (perhaps all but the interest in developing autonomy), but such a coincidence could take place even in a creature who engaged in no critical evaluation of his ends and desires at all. Thus it would seem that acting in accordance with one's objective self-interest is neither necessary nor sufficient for autonomy. If this is true, then the externalist position which makes autonomy depend upon acting in accordance with one's objective self-interest must be rejected.

---

[40] See also the discussion of substantive independence in Haworth 1986, Chapter 1.

The second worry with this externalist requirement is that it is difficult to understand how the externalist conception of real interests is supposed to figure in an account of *autonomy* (as opposed to a theory of pure rationality, say). For autonomy is supposed to be the condition of being *self*-directed, but it is unclear where the self is on such views. The attraction of liberal (internalist) accounts of interests is that they tie interests directly to the real or subjectively defined ideal desires of individuals. That advantage is not available to the externalist. If the emphasis is on autonomy, the absence of a direct connection between the desires and interests of persons becomes even more jarring. For on the externalist conception of autonomy, autonomous action is action which is motivated by a desire to pursue one's objective interests, and autonomous desires must be those which have as their objects those ends which constitute one's objective interests (among other things, perhaps). But one may be completely indifferent, or even hostile, to one's objective interests. In such a case it is difficult to understand how the pursuit of such ends constitutes self-direction.

The externalist could object at this point, claiming that he is relying on a different conception of the "real self" from that which is employed by the internalists. On his view, the self is not to be identified with a person's subjective self-conception, but includes interests which may go beyond her present perception. This is certainly true, but it may not help the externalists' case. For now we can and should demand an account of what this "real self" is. Is it developed or discovered? How does one gain access to it? Do all persons share

some core components of their real selves, those that are identified in the account of objective interests? It is an attraction of the internalist conception of the bi-level theory of desires that it offers an account of how the real or authentic self is developed through a process of self-definition based upon the person's committing herself to some desires, plans and purposes, and rejecting others. It is not clear that the externalist can attain this same advantage, if he insists that the real self can be identified independently of the actual commitments of agents. At least, we need a fuller account of what constitutes the real self on the externalist views considered here.

Finally, there is a pragmatic concern which the externalist thesis raises. Such externalist theories are committed to realism about self-interest. Babbitt presupposes that the deferential wife (and everyone else) has an interest in being autonomous and having self-respect, even though she would reject this interest under subjectively defined ideal choice conditions. As Campbell has pointed out in response to Babbitt, many feminists (and non-feminists as well) object to realism about interests on the grounds that "it appears to require an epistemology that is elitist and undemocratic".[41] Campbell cites Anne Sellers as an example of a theorist who objects to realism about interests on these grounds:

> At best, the use of [realist] epistemology appears to be profoundly undemocratic. At worst, it is an exercise in domination. At best, some women are telling other women what they are like, what their interests are, and how

---

[41] Richmond Campbell, "Comments on 'Feminism and Rational Interests'", unpublished CPA commentary 1992, p. 2.

> they might best be served. At worst, some women are imposing their own interests on the [women's] movement as a whole... How do we know when we are not simply being sold someone else's ideology if we cannot rely on our own judgment?[42]

The problem which Campbell remarks on here is a serious one, though, faced by any externalist position. For in identifying a privileged class of desires as those which all rational or autonomous persons would desire, they leave individuals open to oppressively paternalistic interferences with their freedom, in the name of their objective interests, should they fail to appreciate the conditions of their own well-being adequately. While the justification of such paternalistic interference does not follow immediately from the lack of autonomy, both Babbitt and Benson hold that, since developing autonomy is in the objective interests of individuals, a *prima facie* case can be made for interfering with those who do not desire to develop their critical competence so as to make them realize that they are being personally deprived by the conditions which retard their development of autonomy. More generally, the capacity for autonomy or freedom of the will has served as the basis for many of the libertarian arguments which recognize a strong (even if defeasible) right to be free from paternalistic interference by others. By claiming that individuals whose desires do not comport with their objective interests thereby evidence a lack of autonomy,

---

[42] Anne Sellers, "Realism versus Relativism: Toward a Politically Adequate Epistemology", in Feminist Perspectives in Philosophy eds. M. Griffiths and M. Whitford (Indiana University Press, 1988), p. 172; quoted in Campbell, "Comments on 'Feminism and Rational Interests'", p. 2. Campbell is himself a realist about interests, though, and does not think that such elitism follows from that position.

externalists deny to those individuals an important grounding of the right to be free from such interference. Thus they leave open the possibility of systematic frustration of a person's expressed and considered desires in the name of her real (objective) interests.[43]

IV.4 TOWARD A MORE REASONABLE VIEW

Internalism is not acceptable as it stands. No theory which makes identification with a desire sufficient for its autonomy can avoid the possibility that one might be heteronomous at the level of one's volitions. Nor can it avoid the possibility that there could be autonomous slaves. Some version of externalism must be correct, then. Yet none of the externalist theories which have been offered so far appears acceptable. The internalist perspective allows too much authority to the individual: critical competence here amounts to no more than being able to adopt higher-order attitudes toward one's own desires. This is too weak. The substantive externalists just surveyed, on the other hand, grant too little importance to the actual aspirations and commitments of real agents: critical

---

[43] This was a worry which Berlin took very seriously. See Berlin 1969, pp. 151-152. Benson 1991 offers an externalist account which is similar to Babbitt's. He attempts to answer the charge that such a position allows for coercive intervention to ensure that people pursue their rational, objective interests. His argument is essentially pragmatic, for he argues that the use of brutal or coercive means will not be effective methods for getting people to adopt proper attitudes towards their own interests. This is not satisfactory, however, for it does not answer the charge that a justification could be given if effective means were available.

competence here requires that one be able to recognize one's objective interests and bring one's actual desires in line with them. This is too strong. While one might be able to develop an account of objective interests on the basis of basic needs, these theorists go beyond an uncontroversial account of basic needs. What would a more reasonable conception of the critical competence involved in personal autonomy be?

First, it would reject the implausible internalist commitment to an interpretation of autonomy-conferring identification which is based, ultimately, just on the subjective attitudes of approval or disapproval of agents. The externalist suggestion that individual's have objective interests, specifiable without reference to the actual desires of those individuals, provides one conception of a theory of autonomy which avoids this problematic internalist commitment. But it does so at an unacceptable price. Happily, we need not embrace the externalist commitment to objective interests in order to determine whether or not a desire which is identified with is autonomous. Dworkin provides an alternative theory which can be modified so as to give an acceptable externalist theory of autonomy which does not require that we appeal to the objective interests which all persons share. (Though an account of objective interests need not be cashed out in terms of interests which all persons share universally, this is the sense of objective interests which I believe unites the externalist theories I have been considering

and which I am taking exception to.) I include Dworkin among the externalists[44] because he is aware that approval of a desire, even decisive approval, is not sufficient to confer autonomy upon it. As he remarks, "Authenticity, while necessary for autonomy, is not sufficient. A person's motivational structure may be his, without being his own."[45] This possibility exists because "the identification with his motivations, or the choice of the type of person he wants to be, may have been produced by manipulation, deception, the withholding of relevant information, and so on. It may have been influenced in decisive ways by others in such a fashion that we are not prepared to think of it as his own choice."[46] Dworkin expresses this insight somewhat differently in pointing out that authenticity by itself "leaves no room for false consciousness. An individual may identify or approve of his motivational structure because of an inability to view in a critical and rational manner his situation."[47]

What is needed beyond identification, Dworkin suggests, is that the process of reflection and identification itself be immune from certain sorts of influence: identification must be made under what he terms conditions of "procedural independence" if one's reflective attitudes are to ground autonomy. Thus I shall

---

[44] In doing so I am offering an interpretation of Dworkin's work that is inconsistent with that offered by Christman, who reads Dworkin as providing an internalist theory of autonomy. Interpreted this way, Dworkin's theory is open to the same challenges as Frankfurt's theory. Cf. Christman 1987 and 1991b.

[45] Dworkin 1976, p. 25.

[46] Dworkin 1976, p. 25.

[47] Dworkin 1976, p. 25.

call this view "procedural externalism". Dworkin explains what needs to be done in explicating procedural independence:

> 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. It involves distinguishing those influences such as hypnotic suggestion, manipulation, coercive persuasion, subliminal influence, and so forth, and doing so in a non *ad hoc* fashion. Philosophers interested in the relationships between education and indoctrination, advertising and consumer behaviour, and behaviour control have explored these matters in some detail, but with no finality. [48]

Dworkin's understanding of the task involved in specifying the conditions of procedural independence is found in the following passage.

> The problem of analyzing procedural independence is the task of characterizing those influences which in some way prevent the individual's decisions from being his own.... With respect to autonomy, conceived of as authenticity under conditions of procedural independence, the paradigms of interference are manipulation and deception, and the analytic task is to distinguish these ways of influencing people's higher order judgments from those (education, requirements of logical thinking, provision of role-models) which do not negate procedural independence.[49]

Though Dworkin's condition of procedural independence seems intuitively easy to articulate -- ruling out the inducement of second-order approval through drugs, hypnosis, subliminal suggestion, deception and the like -- it has proved

---

[48] Dworkin 1988, p. 18.

[49] Dworkin 1976, pp. 25-26.

exceedingly difficult to specify in less paradigmatic cases. Does socialization *ipso facto* violate the constraints of procedural independence? Do restrictions on options or freedom of action constitute interferences which violate procedural independence? These are questions which Dworkin's work has left unanswered.

More troubling, Dworkin's own characterization of the task of spelling out the conditions of procedural independence makes it sound as though one must already know under what conditions identification does confer autonomy upon the desires identified with before one can specify the conditions of procedural independence. If that were so, then the account of procedural independence would be redundant. Christman notes this problem with Dworkin's account.

> ... Dworkin claims that acts of identification with LOD's [lower order desires] must "not [themselves be] influenced in ways which make the process of identification in some way alien to the individual." But what *are* the ways that make the act of identification alien to the individual? If we had an account of that, then we would have the key to the autonomy puzzle at every level.[50]

Procedural independence requires that the processes of reflection upon and evaluation of one's desires not be manipulated in various ways. Can the ways of manipulating the reflective process, to some extent at least, be specified independently of the subjective attitudes of the agent, without resort to such objective standards as externalists like Babbitt have imposed, and without making

---

[50] Christman 1987, p. 287.

the conditions of procedural independence redundant or the theory as a whole circular? Surely they can. For our account of critical competence is not completely empty. We know, for example, that critical competence requires that one be able to take one's own desires as objects of evaluation, and so we know that any process which makes one's desires invisible to oneself must undermine the autonomy of those desires. Procedural independence would rule out any forms of influence which impose this sort of barrier to self-knowledge, therefore.

Furthermore, while the evaluation of one's desires will often depend upon other desires one has, there is more involved in critical evaluation than testing one desire against others. The reasons for approving of some desire usually include some beliefs, as well, which can be evaluated against standards other than those provided by the agent's preferences. For a desire may be endorsed simply on the basis of what one believes about one's options, for example, or about what is morally required. While these beliefs may give rise to second-order desires (the desire to act on one's charitable desires, for example), the evaluation and subsequent endorsement of one's desire (to be charitable) may be based on the belief that this is morally required (or morally virtuous, or good), and that belief can be evaluated independently of one's attitude toward it. Such beliefs often figure centrally in the reasons why a person approves of some desires and rejects others. If we want to know whether a person's endorsement of a desire is autonomy-conferring we must, then, examine her reasons for endorsing it; in particular, we will need to determine whether the beliefs which led to that

endorsement are reasonable. Can she provide good reasons for endorsing a particular desire or for rejecting others? This question is central to determining whether identification with a particular desire ought to be considered autonomy-conferring, and it can be answered without adopting the controversial features of the externalist position as it has been articulated.

There are other ways of evaluating an agent's approval of her desires as well. I shall explore some of these in the three chapters to follow. There I shall examine the following three further disputes which divide internalists and externalists:

> (1) Internalists adopt an internalist conception of the rationality of desires. (2) Internalists adopt an internalist conception of values. (3) Internalists maintain that there is a radical separation between questions of freedom of action, on the one hand, and freedom of the will and the autonomy of desires, on the other. Externalists about autonomy reject at least one of the internalist theses, and often more than one.

I shall explain what each of these theses comes to more thoroughly in the next three chapters. Examining these controversies will give us further reason to accept a procedural externalist conception of autonomy and provide us with some means of articulating in a non-circular way what constraints must be satisfied to ensure that identification with a desire is sufficient to confer autonomy upon it.

CHAPTER V

AUTONOMY AND FALSE BELIEFS

V.1 INTRODUCTION

Let us suppose that individuals are socialized and educated in such ways as to permit the development of the cognitive, conative and affective capacities needed for the attainment of agent autonomy. Let us further suppose that individuals do, in fact, critically examine their desires and values, approving of some and so identifying with them, while rejecting others. There still remains a serious problem which needs to be addressed: what if their reasons for making the second-order commitments they do rest on false beliefs?

Clearly beliefs are central to the bi-level theory of autonomy. Given that a second-order desire is adopted because an agent approves of the desire which is its object, we must examine the agent's reasons for that approval.[1] What reasons would induce an agent to endorse a desire? To answer this, surely, we must typically make reference to the beliefs of the agent: beliefs about her options, about the value of what is desired or about the nature of the desire itself, about the instrumental value of what is desired, etc.. To assess one's second-order desires, then, one must evaluate the beliefs which led to their adoption or

---

[1] Those reasons need not, themselves, be construed as "third-order" desires. They might be other first-or-second-order desires which the agent has, or beliefs of a wide variety of kinds.

maintenance. The centrality of beliefs in the evaluation of autonomy is further highlighted when we recall that the general reason for wanting to evaluate one's second-order desires is to ensure that they are themselves autonomous (that is, to ensure that they have not been shaped by illegitimate external influences or modes of influence that violate one's procedural independence). For one of the most effective means of influencing a person's desires is to manipulate her beliefs.

Yet internalists and externalists disagree about whether false beliefs undermine the autonomy of desires which are adopted on the basis of them, or whether having true relevant beliefs should be taken as a necessary condition of being autonomous with respect to those desires to which they are relevant. I will examine this dispute here, and argue that the capacity to form reasonable beliefs is a necessary component of the critical competence which makes autonomy possible. Insofar as the standards by which a belief's reasonableness is determined are external to the subjective psychological states of the agent, the capacity to form reasonable beliefs invokes an external (and hence objective) standard. This is an externalist position, then, but it is weaker than those which have been defended by Richmond Campbell, Susan Wolf and others, who hold not only that one must have the capacity to form reasonable beliefs, but one must exercise it in actually forming reasonable or true beliefs if the desires whose endorsement rests upon those beliefs are to be countenanced as autonomous. I think the weaker thesis is more plausible.

V.2 INTERNALIST AND EXTERNALIST REQUIREMENTS OF TRUTH

Both internalists and externalists recognize that rationality is necessary for autonomy. Yet it is sometimes claimed, somewhat misleadingly, that they disagree over what kind of rationality is needed.[2] In his characterization of critical competence, Haworth argues that its development is intimately connected with the development of rationality. This is not surprising, given our explication of autonomy as requiring that individuals be motivated by authentic reasons for action, which are both explanatorily adequate and defensible. Beginning from the (uncontroversial) premise that "a rational person is one who acts for reasons,"[3] Haworth identifies three "modes of rationality" which, he argues, are each necessary for the attainment of agent autonomy (or what he terms "normal autonomy"). These different "modes of rationality can be distinguished by the sorts of matters the rational person is required to have reasons for."[4] As we shall see, these modes of rationality can each be given an objective or subjective characterization; if they are taken to be necessary for autonomy, it will matter which characterization we adopt.

First, autonomy requires "technical" or "instrumental" rationality. The "*technically* rational person has reasons for adopting the means by which he

---

[2] Cf. Christman 1991, pp. 349-350; Lindley 1986, Part 1; Susan Wolf, Freedom Within Reason (Oxford University Press, 1990), Chapter 4.

[3] Haworth 1986, p. 27.

[4] Haworth 1986, p. 27.

pursues his ends, but not for his ends."[5] Second, autonomy requires "economic" rationality as well. As Haworth characterizes economic rationality, "the *economically* rational person chooses rationally from among a given order of ends (preferences for outcomes), but he has no reasons for ordering them in the particular way he does."[6] Finally, Haworth identifies a condition which he calls "full rationality": the *fully* rational person has reasons for his ends themselves, that is, for the ordering he has imposed upon his desires.[7]

Using this schema, we can see that the debate between internalists and externalists does not centre on the kind of rationality that autonomy requires, for both can agree that autonomous agents must have reasons for their desires in all the senses here identified. They can give reasons for choosing certain means to their ends, for choosing certain options, given the priorities they have among their preferences for outcomes, and for having the ends and priorities they do. This final condition of rat nality, requiring as it does that one have reasons for one's desires and ends, would be controversial if it required a rejection of the general Humean thesis that desires are basic and cannot themselves be assessed in terms of their rationality or irrationality. But the point here is a different one: an autonomous agent has reasons for endorsing some of her desires as reasons for action and rejecting others. This is compatible with a Humean psychology.

---

[5] Haworth 1986, p. 27.

[6] Haworth 1986, p. 27.

[7] Haworth 1986, p. 27.

Certainly full rationality is required for autonomy on the bi-level theory, for one must have reasons for approving of one's desires and ends to be autonomous with respect to them.[8]

The crux of the disagreement between internalists and externalists, then, is not over whether agents must have reasons for what they desire and what they do, but, rather, concerns whether these reasons must be defensible or true or correct by external standards.

Consider, by way of illustration, the condition of instrumental rationality. The question on which internalists and externalists divide is not whether this is necessary, but whether what one instrumentally desires must be sufficient for gaining its end. This, in turn, will often be determined by the truth or falsity of the beliefs upon which one's instrumental desires are based. For one can be technically rational, i.e., be able to give reasons for one's conditional desires, yet choose means which are insufficient for gaining one's end because one has relevant false beliefs. The following example illustrates this possibility. A young man, Joe, wants very much to impress a woman of his acquaintance, Anne. We can suppose that this desire conflicts with no others that he has and that he wholeheartedly identifies with it. He also believes that the best way to achieve his end is by demonstrating his superior courage and daring. Accordingly, he seeks out opportunities which involve risks to his security as a means of displaying his bravery. But Anne does not value excessive courage, and thinks

---

[8] Cf. I.1.

Joe is reckless in his behaviour. Here Joe has chosen means which are insufficient for gaining his end, as a result of faulty instrumental reasoning (faulty because the relevant belief is false).

From an internalist perspective, Joe's behaviour is both rational and autonomous. Given the desires and beliefs he actually has, he displays instrumental rationality in seeking out dangerous situations. His desire to do so, then, must be considered autonomous; it is such that he could offer good reasons, from his own perspective, for having and endorsing it. But those reasons depend on false beliefs. The externalist would not agree with the internalist's judgment concerning the autonomy of Joe's instrumental desire or action. For the externalist would insist that instrumental desires satisfy the requirements of technical rationality only if they are based upon true beliefs (or beliefs which are adequately supported by evidence) and are sufficient for gaining the end to which they are instrumental.

The same sorts of considerations hold with respect to economic and full rationality. Given that beliefs are relevant to one's practical deliberations about how to rank one's ends, as well as what ends and desires one ought to adopt or retain, there is a possibility of error in one's practical reasoning. The internalist will insist that an end is autonomous just in case one's other desires and beliefs give one a reason for having it and endorsing the desire for it as a reason for action. (This is tied to their general rejection of a "realist epistemology": for most internalists, a person satisfies the requirements of rationality provided that her

actual beliefs and desires cohere together in the right way.) The externalist, by contrast, will require that one's other desires and beliefs be "justified" by external standards of reasonableness, evidence or truth. (This implies, in turn, a realist epistemology.[9]) The ultimate point of disagreement between them, then, hinges on the question of what evidence or truth conditions ought to be imposed upon the beliefs which are relevant in one's practical deliberations about what one desires and one's desires themselves.

Internalist analyses of autonomy which employ agent approval as the basis of desire autonomy (whether it is approval of one's occurrent desires or the processes by which they have been formed) have largely ignored the role of beliefs in motivation. Thus Frankfurt does not offer any systematic treatment of the concern that having false beliefs could undermine the autonomy of one's desires. Christman has explicitly discussed the internalist's position on beliefs and rationality, however, and so I will concentrate on his theory.

In "Autonomy and Personal History", Christman claims, in effect, that what he says about the autonomy of desires can be extended *salva veritatis* to beliefs. Thus he says

> I think that my general account of autonomy can be applied to belief formation more or less without alteration. One is autonomous if one comes to have one's desires *and* beliefs in a manner which one

---

[9] Cf. Section IV.3. It should be noted that most of the authors considered here (except for Wolf 1990) do not explicitly discuss such metaethical and epistemological issues explicitly, and so I offer this suggestion only tentatively.

> accepts. If one desires a state of affairs by virtue of a belief which is not only false but is the result of distorted information given to one by some conniving manipulator, one is not autonomous just in case one views such conditions of belief formation as unacceptable... All that I would reject in this vein is the view that one lacks autonomy *simply* because one's beliefs are false.[10]

Christman offers virtually no argument for his position here.

What Christman's position comes to, though, is that a belief is fully attributable to an agent (autonomous?) so long as the agent did not (or would not have) resisted its formation if he were fully informed as to the processes of its formation, he were minimally rational and the beliefs so adopted were not manifestly inconsistent.

This is, of course, too strong in one sense, because it is irrational for an agent to believe that his beliefs are even manifestly consistent.[11] Any stipulation that an agent's beliefs must not be inconsistent is too strong. For a rational agent must have inconsistent beliefs. As Campbell (and Gilbert Harman) have argued, a rational agent will know that he, being fallible, will have at least one false belief. Hence the total set of his beliefs cannot all be true together, i.e., the total set will be inconsistent. It is possible, of course, that a person might not know that his beliefs must be inconsistent in this sense (and so not recognize the inconsistency

---

[10] Christman 1991b, p. 16.

[11] See Richmond Campbell's, "Can Inconsistency be Reasonable?", Canadian Journal of Philosophy Vol. XI, No. 2 (June 1981). It is also too strong as a condition for autonomous desires, since it is equally unreasonable to suppose that all of one's desires are consistent (or mutually satisfiable).

as "manifest"); but then he fails to be rational. Thus the two conditions of minimal rationality and consistency of beliefs are mutually unsatisfiable.

But Christman's historical condition is too weak to ensure the truth of one's beliefs, for one could have full information concerning the processes of one's belief-formation and yet still have false beliefs.[12] The falsity of some beliefs can only be discovered with future experience, for example, and so they could pass the historical test yet still be false. Moreover, a belief could be formed through only legitimate processes, be supported by strong evidence and challenged by no counter-evidence, yet still be false.

In insisting that the judgments which go into the assessment of one's desires and the processes of their formation must be minimally rational for the reflective agent, Christman is giving voice to one possible internalist position concerning rationality. Christman begins (following Brandt[13]) by distinguishing between "internalist" or "subjective" accounts of rationality and "externalist" or "objective" accounts. By an internalist account of rationality, Christman means that rationality is determined only by the beliefs and desires which the agent actually has, those that are "internal" to the agent. More specifically, Christman defends an internalist view of rationality which demands only "that the beliefs (upon which the person's conditional desires are based) are consistent and the desires

---

[12] For an elaboration of this criticism see Benson 1991 and Babbitt 1993.

[13] Richard Brandt, A Theory of the Good and the Right (Oxford, Clarendon Press, 1979).

(whether conditional or "brute") are transitive."[14] Clearly this is a very minimalist account of rationality. It does not insist that the beliefs upon which one's conditional desires depend be well-founded by any objective standard, for example:

> On an internalist account, the property by which an action is considered rational for an agent bears only on those beliefs and desires actually "internal" to the agent, not on the relation between those beliefs and the world (i.e., a relation of fit or accuracy).[15]

Nor does Christman accept even the rigorous internalist standards of rationality required by most theorists who define the economic rationality of desires in terms of utility-maximization, where preferences must be ranked along an ordering that is not only transitive but complete and continuous as well, meaning that a single ranking will be compiled for all available objects of preference.

Christman worries that imposing an "evidence requirement" of external rationality for beliefs will make autonomy indeterminate. Will desires which are based on "better" evidence be freer or more autonomous than those which are based on less adequate evidence? If we employ a threshold criterion for evidence we run the risk of adding to the problem of degrees above the threshold a further

---

[14] Christman 1991, p. 350. Thus Christman demands only very weak forms of technical and economic rationality. Insofar as he defends a modified version of the bi-level theory of desires, however, and argues that the ability to become aware of and critically evaluate one's desires and the processes by which they were acquired is necessary for their autonomy, he is also committed to the necessity of full rationality, in Haworth's sense.

[15] Christman 1991, pp. 149-350.

problem of arbitrariness in demarcating that crucial point.[16] These are not decisive objections, of course, but they do signal a need for care. Unless one can produce serious reasons for imposing such a difficult condition, then, one ought to refrain.

We must be careful to distinguish Christman's position from another, with which it might be confused here. It is surely true that a person need not have true beliefs alone (even true relevant beliefs) to be an autonomous agent. To insist on such a rigorous requirement would, of course, make agent autonomy impossible for fallible creatures such as human beings to attain. But his position is not this uncontroversial one. For he is claiming that the truth or falsity, reasonableness or unreasonableness, of one's beliefs is, in itself, irrelevant to the question of whether or not the desires which one approves of, even on the basis of those beliefs, are autonomous.

Understood in this light, in relation to questions of desire autonomy, the internalist position seems to be open to decisive counter-examples, however. Consider, by way of illustration, Paul Benson's treatment of the role that false beliefs play in the socialization of women to adopt oppressive norms of feminine appearance. He notes that women are socialized to accept many false beliefs, which reinforce their commitment to meeting these norms of attractive or acceptable physical appearance. Such beliefs include the following: that women's physical appearance is naturally defective and so must be "fixed", that meeting

---

[16] Christman 1991, p. 356.

social standards of beauty is necessary for one's social success, physical and mental health, and personal worth.[17] These beliefs are widely held, and the ideals of feminine appearance are systematically inculcated and reinforced. Insofar as the socialization of women into the ideals of feminine beauty is effective, women will desire to meet these ideals. The beliefs mentioned above will reinforce that identification, moreover, and provide a basis for (give them reasons for) identifying with those desires, thus making it very difficult (at least) for women to acquire a more adequate understanding of the place which physical appearance should occupy in their self-conception and sense of personal worth. Insofar as women adopt and internalize the ideals of feminine beauty only because (or largely because) they have acquired such false beliefs, it is difficult to conclude that their identification with those ideals confers autonomy upon them or that they are truly acting in a self-directed manner when they mutilate, starve, pluck, wax, shave, paint, polish, pierce and adorn themselves in whatever manner is dictated by current fashion.

What is particularly insidious about such cases of oppressive socialization which are reinforced by the inculcation of false beliefs is that the persons who have been socialized in these ways and who have adopted such false beliefs could come to know how they developed their desires and beliefs yet continue to approve of them. If women are persuaded that their well-being so firmly depends upon meeting the standards of physical attractiveness approved of

---

[17] Benson 1991.

within their society, then even coming to know that their beliefs have been forced upon them may not persuade them to reject them. They may see it as, at worst, a necessary evil: the lessons had to be learned, after all. Or they may even be grateful that they were subjected to such experiences: given that their well-being depends upon their having been taught the proper standards, those who have taken the time and trouble to see that they learned and internalized the proper ideals have done them a great service. In these cases, Christman's historical test would fail to offer any grounds from which we could deny that the desires which are approved of on the basis of these beliefs are fully autonomous. Yet, intuitively at least, such desires seem heteronomous.

In light of such difficulties with the internalist view, externalists offer more rigorous conditions of rationality and the justification of beliefs than are accepted by internalists, and it is to their views that I now turn. As Christman characterizes this distinction between internalist and externalist conceptions of rationality,

> The internalist would demand only that a person act for *reasons* (perhaps ones which meet some requirement of consistency), while the externalist demands that the free agent must act in accordance with *reason*, where that includes knowledge of the truth, both about the world as well as morality.[18]

An externalist account of rationality might thus impose a number of further conditions concerning the rationality of one's desires and beliefs. Most would

---

[18] Christman 1991, p. 350.

insist that one's beliefs reasonably or even accurately represent the world, that they be based on good reasons or objectively adequate evidence, that they be based on relevant information, etc.. Some, but not all, externalists would also insist that one have objectively adequate moral beliefs, i.e., justified or correct values. (I shall postpone the discussion of normative beliefs until the next chapter.) Other externalists insist that the beliefs upon which one's conditional desires depend, and one's beliefs about probabilities concerning utilities, must be true. Haworth's own characterization of technical, economic and full rationality is an externalist account, as I am using that term. Richmond Campbell and Susan Wolf have also offered externalist conceptions of rationality.

Haworth's understanding of the demands of rationality, together with his claim that rationality is necessary for autonomy, provides an externalist conception of autonomy. For it is not sufficient, on his view, that agents have reasons for choosing some means to their ends rather than others, or that their preference-ordering provides them with reasons for choosing one option over another, or even that they have reasons for imposing the ranking on their preferences that they have; those reasons must be "good reasons", and what determines whether a reason is a good reason is settled, at least in part, independently of the subjective states (beliefs and desires) of the agent whose reasons they are. Thus, to be technically rational, for Haworth, requires not just that one adopt means to one's ends, but that those means be efficient and effective. Likewise, to be economically rational, one must actually choose that

action which would be selected by whatever decision rule one is employing to determine rational choice (e.g., the rule of maximizing expected utility). It is not enough that one have reasons for selecting one option rather than another, in other words: one must have the right reason, i.e., that one's choice is justified by the decision rule. Finally, in developing full rationality one must subject one's preferences themselves, and their ordering, to critical scrutiny. The end of such reflection is an answer to the question whether one wants to retain a particular desire or preference as one's own, with the same ranking that it actually has. The sorts of questions that Haworth mentions as relevant to this assessment are (1) whether the preference is "accurate", i.e., whether acting on it would satisfy the agent; (2) whether acting on the preference would bring about consequences that, had they been foreseen, would have prompted the person to revise it; (3) whether the preference is founded on "opinions" (beliefs) that the person has good reasons for holding; (4) whether the preference is consistent with principles the person holds; (5) whether the preference is consistent with values to which the person is committed. These considerations all bear on the question whether one has objectively good reasons to order one's preferences in the way one does, or to revise that ordering. If reflecting on one's preferences according to these considerations leads one to conclude that there are good reasons for one's present preference-ordering, then one has made that ordering fully one's own. Provided that the reasons one can give in favour of one's ordering are (a) relevant, (b) sufficient and (c) acceptable (by external criteria of truth or reasonableness), then

one's commitment to that ordering is fully rational, and one acts autonomously when one acts in accordance with it. One can give good reasons, not only for doing what one does, but for wanting what one wants.[19]

It is because Haworth imposes a condition of acceptability on one's relevant beliefs, insisting that they be well-founded, that I place him in the externalists' camp. Haworth does not discuss this condition very fully, however, or explain exactly why he takes ill-founded beliefs to pose a threat to the autonomy of one's desires, and so I shall turn now to the writings of Campbell, as he takes up these questions directly. Both an absence of true and relevant beliefs (ignorance) and the presence of false beliefs (cognitive error) can influence the formation or maintenance of second-order volitions. Campbell calls those desires which are sustained only by false beliefs "false desires".[20] We may isolate a number of general types of false beliefs here: mistakes about the world, mistakes of instrumental rationality, and mistakes about the nature and value of one's desires themselves.

It is clear that false factual beliefs can lead to the formation of a false desire, but what we must be concerned with is the role of false beliefs in the formation of second-order volitions, that is, false beliefs which lead reflective agents to approve of and identify with some of their desires and to reject others. Consider the case of Bob as an illustration of the general way in which false factual beliefs

---

[19] Haworth 1986, Chapter 2.

[20] Campbell 1979, pp. 181-192.

can lead to the formation of a false desire. Bob is a kind person. He believes it is morally correct to help others who need assistance, at least when such aid can be rendered without great cost to oneself. He also believes that Eric is in trouble and needs assistance, and that he can render effective aid. Reasoning on these beliefs, and being disposed to help those in need, Bob desires, at the first-order, to render aid to Eric. But suppose that Bob's belief that Eric is in trouble is false. The desire to which this belief is relevant (the desire to assist Eric) must also be false, where a false desire is one that the agent would cease to want to be moved by if the falsity of the belief sustaining it were to be revealed and appreciated by the agent.

Bob's situation, at least as stated, does not show that any of his second-order desires are false, however. By characterizing Bob as a kind person we might suppose that he has desires to be kind toward others, and that he approves of these desires at the second-order. His general altruistic volition would not be undermined by the discovery of his cognitive error concerning the needs of Eric. If Bob had reflected upon and endorsed at the second-order his specific desire to aid Eric, then this volition (i.e., the volition to act on his desire to help Eric) would have been false. According to Campbell, insofar as the desire to assist Eric and the volition to act upon this desire are developed and sustained only on the basis of a false belief, neither the desire nor volition is autonomous.[21] For, it will be recalled, Campbell characterizes autonomous desires (i.e., self-identifying

[21] Campbell 1979, pp. 105-207.

desires, in Campbell's terminology) as those one could adopt and sustain if one were fully informed of and sufficiently attentive to the relevant facts concerning one's desires and their objects. Those desires that are sustained only on the basis of false beliefs obviously could not be sustained under conditions of full, relevant information.

A more serious case of a false volition could be the result of false factual beliefs about one's options. Suppose a person, Jerry say, has become convinced that he has only one career option -- to follow his father in the family fishing business -- and that this belief is false. It may be that Jerry has come to this belief because he has other false beliefs (about his own abilities to succeed in alternative career choices, for example); his cognitive error may have been caused by the manipulation of information by others or not. If Jerry has desires to pursue careers other than fishing, there will be a tension between his desires and his belief about his options. This dissonance can be relieved by his rejection of these alternatives and his endorsement of his desire to fish.[22]

To determine whether Jerry's approval of his desire to fish is false or not, we would have to know whether the correction of the false belief about his options would lead to a corresponding change in his volition to act on his desire to fish.

---

[22] This may be a case of "sour grapes" adaptive preference formation, and so may pose additional problems for determining whether it is an autonomous choice. Cf. Jon Elster, "Sour Grapes -- Utilitarianism and the Genesis of Wants", Utilitarianism and Beyond eds. A. Sen and B. William (Cambridge University Press, 1982). The issue of adaptive preference change is taken up in Chapter VII, and so I will not consider this complication further here.

Is the volition sustained only on the basis of his false beliefs? If so, then both the volition and its object are false desires in Campbell's usage. As such, they are not autonomous desires.

A different kind of cognitive defect is involved when one makes mistakes of instrumental rationality. We have already seen an instance of this kind of defect in our example concerning Joe and Anne. It will be recalled that Joe has a first-order desire to impress Anne but has chosen means which are insufficient for achieving this end due to a false belief concerning what will impress her. While neither his desire to impress Anne nor his volition to make that desire effective is false, his instrumental desire to seek out and withstand dangerous situations, as a means to this end, is false. He would revise this desire if his false instrumental belief were to be corrected. Thus, while his desire to impress Anne and the volition to act on that desire may be autonomous, the desire to confront dangers is not.

I take it that the kind of error which Joe falls into is not uncommon. But, insofar as such errors typically involve the formation of only first-order desires, instrumental mistakes of this kind do not pose a serious challenge to the formation of authentic volitions by agents. While autonomous agents must care both about what they do and why they do it, such concern surely manifests itself typically in an evaluation of one's important projects, plans and values: one's instrumental desires will not often be taken as objects of critical evaluation, unless one becomes aware of the possibility that they are insufficient for attaining their

end. Even if he should come to such awareness, however, revising one's instrumental desires need not produce any alteration of one's self-identifying projects or volitions.

Campbell's discussion of false desires is an important contribution to the debate. It is clearly an externalist account, for Joe, Bob, and Jerry all have reasons for the desires that they have, yet it is claimed that, because those reasons are false from an external perspective, they are not autonomous desires. On this view, a desire is false only if an agent's acquiring the relevant true beliefs would change that desire. Both the desire and the second-order desire to gratify or resist it must be susceptible to change if it is a false desire.[23]

Campbell's analysis of false desires gives a reason for thinking that such desires cannot be truly autonomous. For what is common to false desires is not only that they would be revised if the error on which their formation depends were to be exposed, but that the agent's desires themselves would provide good and sufficient reason to reject them if he had true relevant beliefs. If the belief changes, then the agent will have, in virtue of his own motives, sufficient reason to change his preferences and his higher-order attitudes towards his preferences. Thus, we must suppose that Bob has other desires and beliefs which are not themselves false (a desire not to waste his limited time and resources helping those who are not in need of assistance, that it would be intrusive to help those not in need, etc.) which would provide him with good and sufficient reason to

---

[23] Campbell 1976, p. 181.

alter his desire to help Eric upon discovering that he is not in need of aid. And of Jerry we might suppose that his rejected career desires, if they were stronger than his desire to fish, would displace the false desire to fish once he comes to recognize that they are viable options. Joe, likewise, would have sufficient reason, given his actual motives, to change the desire to face danger if he had the appropriate true beliefs. If the correction of a false belief would lead to a corresponding change in an agent's desires, and that change would be brought about because of the real motivations of the agent, then, in the absence of that corrected belief, the false desire cannot be fully attributed to him. Coming to have true and relevant beliefs would lead the agent himself to reject the false desire as a reason for action.

There are problems with the externalist's insistence upon the truth of one's beliefs (even relevant beliefs), however. We cannot simply assume that all desires and values which are based on or supported by false beliefs are non-autonomous. Campbell's position is more reasonable than this, for he claims that a desire is non-autonomous only if its sole support is a false belief, that is, if it is acquired or maintained only because of ignorance or a cognitive error. Yet even this will not do. For individuals sometimes remain wilfully ignorant of relevant facts so as to maintain certain self-identifying values or desires (racist and sexist examples come readily to mind here); though such desires are sustained only because the individual maintains some false beliefs, we might nonetheless want to say that those desires and values are fully attributable to the agent who maintains and

endorses them. Alternatively, a person might have certain desires and endorse them because they are supported by false beliefs, yet even she could acknowledge (before and after disclosure of the relevant facts) that the desire was truly hers. Suppose, for example, that a person has a desire to be a professional pianist. Upon reflection she might wholeheartedly endorse that desire. Yet we might also imagine that that endorsement is based on a number of false beliefs: she believes, contrary to fact, that she has outstanding talent on the piano, that she will enjoy riches and fame, that a pianist's life is one of glamour and ease. With greater experience she may come to realize that these beliefs are false, and when she does she may abandon her desire to be a pianist, identifying with it no longer. Yet she could still recognize that that desire had been hers, that it represented an important part of who she was at the earlier time. It is not obvious, at least, that we must conclude that it was not "really" her true desire just because she will change it when she realizes that her approval of it was based on certain false beliefs.

Susan Wolf has also argued that critical competence requires an externalist conception of the rationality of beliefs. To be fully free one must be able to cognitively "recognize and appreciate the world for what it is."[24] In Freedom Within Reason, she argues that one's freedom is a function of one's ability to recognize and appreciate whatever objective reasons there are in favour of and

---

[24] Wolf 1987, p. 145.

against the various alternatives one faces.[25] One must not only have this ability, furthermore, but one must exercise it in determining which action or choice would be right, and act according to that judgment.[26] Thus, one must have not only the ability to recognize and appreciate "the True and the Good", but that appreciation must determine one's will and action. One's beliefs must be "controlled ... by perceptions and sound reasoning that produce an accurate conception of the world rather than by blind or distorted forms of response"[27], and one's desires and beliefs and actions must conform to the truth so recognized.

The externalist conception of rationality, as set forth both by Haworth and Wolf, has much to recommend it as a necessary condition of the critical competence that is needed to be self-directed. Haworth's position is compelling if one concentrates only on the development of the *capacity* for rationality, in the senses he identifies. That one be able to recognize the reasons there are for and against the ordering which one has imposed upon one's desires, as well as the ability to make rational choices based upon that ordering and to choose effective means to one's self-chosen ends, are all necessary for agent autonomy. Likewise,

---

[25] Wolf 1990, p. 142.

[26] If Reason does not uniquely support a single option, then one must choose from among the set, the members of which are all supported by sufficient reasons to show that they are right.

[27] Wolf 1987, p. 145. One might be concerned here that this establishes a false dichotomy; for surely there are degrees of evidence or standards of reasonableness which fall short of giving one an accurate conception of the world without implying that one has formed one's beliefs in a blind or distorted way.

Wolf's position is extremely plausible if one concentrates on the *capacity* to form true relevant beliefs and to appreciate the objective reasons there are in favour of and against alternatives from which one must make choices. If one lacks this ability over a large number of beliefs, one would be incapable of agent autonomy, for such a condition would undermine the critical competency necessary for the various kinds of rationality Haworth has identified. In less extreme cases, where the inability to recognize and appreciate the reasons which are relevant to one's options is not very widespread, but is confined to a range of specific desires and beliefs, lacking the capacity to appreciate the reasons there are will undermine the autonomy of one's specific desires. The person who suffers from paranoid delusions is not autonomous with respect to those desires which result from his paranoid beliefs, because, at least with respect to them, he lacks the ability to form true beliefs. Such a person has many false beliefs, and even if he reasons in conformity with all the canons of practical rationality from them, many of his desires will be false (those of his motivations which are sustained only on the basis of these cognitive errors). The paranoid seems not to be autonomous with respect to his false desires.

Haworth's position is more problematic if we take it to mean not just that a certain capacity is presupposed by autonomy, but that a desire is autonomous only if it has been adopted in conformity with those standards of rationality, given only true beliefs. Wolf's view, too, is much less plausible if we take the actual condition of having all of one's beliefs and desires determined by the

recognition and appreciation of the True and the Good (i.e., only true beliefs) as necessary for agent autonomy. Indeed, if this were necessary for agent autonomy, then no one would be autonomous. It is even too strong as a condition of desire autonomy. Someone may have the capacity to form true beliefs yet fail to do so, because of a kind of intellectual laziness, for example. Must we conclude that this person's desires are non-autonomous if they rest on false beliefs? There seems no obvious reason why we must. Or consider the following. Suppose a group of friends plan to go sailing, but their plan is based upon the false belief that the weather will be fine. If some reflection on this plan has been triggered (by a debate about this proposal compared with other possibilities, perhaps), they might come to explicitly endorse it. We can suppose that they satisfy the requirements of agent autonomy, and under these conditions the desire to go sailing seems equally autonomous. If so, then they act autonomously when they attempt to carry out their plan, though they will fail because it is based upon a false belief and would be revised if they came to have relevant true beliefs. They seem neither directed by another nor inner impelled by unreflective passion or inner compulsion; if, as Haworth believes, the only possibilities are that one is self-directed, other-directed or inner impelled, the individuals under consideration seem self-directed. Desire autonomy does not require that the desires agents endorse be supported only by true beliefs. Nor does desire autonomy require that agents have full information -- about the nature of their desires and the reasons why they have them and approve of them, or about all the objective

reasons that are relevant to their assessment. Agent autonomy is a condition which allows imperfect agents to direct themselves in what they do in an imperfect world. In exercising their autonomy, they will make mistakes, they will fall into error, and they will need to revise their commitments in the light of new information and changing beliefs in the process of living an authentic life.

V.3 A MORE REASONABLE VIEW

Is there a principled way of distinguishing those cases of false desires which are nonetheless autonomous from those which are not? The reasons which support false desires and identification with them are inadequate for that task, but our understanding of this is given in counter-factual terms: the agent would have good and sufficient reason to revise his desires and withdraw his attitude of approval from those which are false if he had true and relevant factual, instrumental and normative beliefs. But, insofar as the agent lacks such true beliefs, he could claim a false desire wholeheartedly as his own. In doing so, moreover, he could cite what he takes to be good and sufficient reasons for wanting to act on those desires. How can we gain any purchase for the claim that the desire is not really his, then, or that his second-order desire to gratify it is non-autonomous? To be directed by false desires may still be self-direction.

I shall argue that we need to know the answers to two questions in order to determine whether a desire can be attributed to an agent, even when the agent's

identification with it is based on false beliefs: (1) Was the person responsive to the reasons in favour of and against his beliefs and desires? (2) Could the person have avoided acquiring the false belief? I shall explain what I take the relevance of these questions to be here. The conclusion will be that a false desire may be autonomous so long as the person whose desire it is was "reasons-responsive" and his ignorance or cognitive error was not unavoidable. The first of these conditions takes the psychological capacity to form respond appropriately to the (objective) reasons there are both for and against ones desires and values as necessary for autonomy, while the second recognizes that one's external circumstances can sometimes interfere with this capacity without directly undermining the psychological skills which constitute it.

(1) Reasons-Responsiveness[28]

To have and display the general cognitive, conative and affective capacities required for agent autonomy, an individual must be responsive to various sorts of reasons in the evaluation of her desires, values and beliefs. Her conative states, and the evaluation of her first-order desires in particular, ought to be responsive to reasons provided by her affective states, for example. Experiences of pleasure, satisfaction, self-esteem and the like all provide reasons in favour of gratifying the

---

[28] I take this term from John Martin Fisher, "Responsiveness and Moral Responsibility", Responsibility, Character, and the Emotions ed. Ferdinand Schoeman (Cambridge University Press, 1987).

desires which give rise to such feelings, and for approving of them as part of oneself. Conversely, if one experiences dissatisfaction, regret or shame upon gratifying some desire, these feelings provide reasons against one's continued endorsement of it. One ought to be responsive to reasons provided by one's affective states in the evaluation of one's desires. The same is obviously true with respect to one's cognitive states: one's beliefs often provide reasons for or against the endorsement of a given desire, and an autonomous agent must be responsive to such reasons.[29] I do not mean to imply that we can mark a sharp divide between these various facets of our mental life, of course, or that there must be complete harmony among them for the attainment of agent autonomy. I am claiming only that one must be sensitive to the input of all the kinds of reasons one's own states provide in the assessment of one's desires, beliefs and values.

But reasons-responsiveness requires more than this. Autonomous persons must also be responsive to the external reasons which are relevant to their beliefs in particular; they must be responsive to the evidence in favour of or contrary to their beliefs, we might say. The person suffering from paranoid delusions whom we considered earlier not only has many false beliefs, but her beliefs are not responsive to the evidence against them. All counter-evidence is reinterpreted so as to support the paranoid's beliefs, or it is repressed. Even if she is presented with good and sufficient reasons to reject her false beliefs, she will persist in

---

[29] Haworth describes a similar condition of autonomy as a "sensitivity to feedback". Cf. Haworth 1986, Chapter 2.

holding them. Such an agent is not reasons-responsive in the sense meant here.[30]

Reasons-responsiveness requires, then, that a person be responsive to relevant evidence for and against her beliefs. This applies equally to beliefs about the world, instrumental beliefs, normative beliefs and beliefs about herself. A person who is incapable of responding to reasons in any of these areas would be incapable of rational agency under any interesting description. Suppose a person were incapable of forming reasonable beliefs about the world, i.e., were incapable of acquiring reasonable beliefs about the circumstances in which she must make decisions and act. Such a person might have unreasonable beliefs about the options which are open to her, about the resources which are available to her or which are required to carry out her plans, about the desires of others, etc.. Such unreasonable beliefs would be very likely to undermine agent autonomy altogether. To maintain such beliefs would be evidence that the person's beliefs are not reasons-responsive, as opposed to just being false.

I want to concentrate here on a person who is incapable of acquiring reasonable beliefs about herself. Imagine a person who is incapable of forming reasonable beliefs about herself. This might be due to various forms of mental illness, delusions, neuroses, senile dementia or whatever. But these would be

---

[30] This condition is very close to the ability to recognize and appreciate whatever relevant objective reasons there are, as Wolf discusses it. But it is only having the ability, and not its exercise in the formation of reasonable beliefs, that I am claiming is necessary for agent autonomy and for the autonomy of particular desires.

extreme cases. There are more mundane examples of the phenomenon I have in mind.

The most obvious cases of the inability to acquire true beliefs about oneself are displayed by those who are self-deceived about their own desires. Self-deception clearly threatens autonomy, for if one is deceived about one's reasons for action then one cannot subject them to critical examination.[31] There are other ways in which a person might be incapable of forming reasonable beliefs about herself which likewise make the attainment of personal autonomy difficult. A person who cannot form reasonable beliefs about what will make her happy or about her own talents will no more be capable of self-direction than the person who is incapable of forming reasonable beliefs about what motivates her to do what she does. Someone who has unreasonable beliefs about the means to her own happiness or concerning her own talents can be expected to experience chronic frustration, regret and dissatisfaction. Though we need not suppose that autonomous agents singlemindedly seek satisfaction or pleasure in directing themselves, such conditions as chronic frustration seem to preclude the conclusion that the person who suffers thus is living autonomously.

This account of reasons-responsiveness can be used to make clearer how false

---

[31] That one not be self-deceived about one's own motivation (i.e., that one have true beliefs about one's own conative states) is a condition which both Dworkin and Christman accept as necessary for autonomy with respect to those states. Thus they would admit that, at least in this one case, having false beliefs undermines the autonomy of one's desires. Cf. Dworkin 1976, p. 26 and Christman 1991b, pp. 16-17.

beliefs can undermine the autonomy of the desires based upon them: if a particular desire, and approval of it, is sustained only by a false belief (or set of false beliefs) which is not reasons-responsive, then that desire is not autonomous, even if the individual wholeheartedly identifies with it. A person who comes to believe, through fierce indoctrination, that another individual is the messiah returned to earth, may form and identify with numerous desires on the basis of such a false belief: the desire to serve him, to spread his message, etc.. If the belief upon which these desires depend is not responsive to reasons against it -- evidence that the "messiah" acts sinfully, that he is mortal, and the like -- then the desires are not only false, but non-autonomous as well.

To be capable of self-direction requires that one be able to form reasonable beliefs about the world, about oneself and about morality.[32] It requires, furthermore, that those beliefs play a proper role in one's practical reasoning. Nothing in the account of reasons-responsiveness implies that those who are responsive in this sense will form only true or even reasonable beliefs, however. While I take it to be obvious that one of the most reliable means of acquiring reasonable beliefs is to develop the skills which characterize reasons-responsiveness and to be diligent in exercising them, one could be responsive to the reasons there are either for or against one's beliefs and desires and yet form

---

[32] The ability to form reasonable, as opposed to true, beliefs may be all that Haworth has in mind when he insists that one's beliefs must be "acceptable". This interpretation is supported by much of what he says in Chapter II concerning "good reasons". Cf Haworth 1986. If so, my view is closer to his than to that offered by Wolf and Campbell.

unreasonable beliefs and false desires. Reasons-responsiveness involves the capacity to respond appropriately to evidence. Reasonableness is an objective feature of one's beliefs and desires. A person could be reasons-responsive without having reasonable beliefs. He could have the capacity to respond appropriately to evidence, but not do so, with the result that his desires and beliefs are not reasonable. Furthermore, a person could have reasonable beliefs and desires without being reasons-responsive. It will be a happy accident that he does have reasonable beliefs (a matter of moral luck), owing perhaps to fortuitous parental influences, in the absence of the ability to respond appropriately to evidence; but such a case is possible. Thus the capacity to respond to reasons and the actual formation of reasonable beliefs and desires are independent conditions. I am arguing that only the capacity to respond to reasons is necessary for agent and desire autonomy.

Being responsive to the reasons there that are relevant to one's beliefs about one's desires and talents need not produce true or even reasonable beliefs. One could be responsive to the kinds of reasons which are relevant to this species of self-evaluation yet still fall into error. This may be due to inexperience, either with what one wants or with the other options which one discounted too readily because of their unfamiliarity. Upon further experience, a person who has been pursuing a desire, a desire to be a professional artist, say, which is false in this way, may come to realize that she does not really want what she thought she wanted. She will, then, have reason to revise her desires. A reasonable

description of this might be for her to say, "I wanted to pursue my art back then, but now I know that what I really want to do is teach." Even after the revision of one's preferences, one could still acknowledge that one's former desires were really one's own, though that is no longer true. Insofar as one's former desires were based on some reasons, one could now acknowledge that they had been ~. e's own without experiencing guilt, shame or loss of self-respect. This may not be true of false desires which are sustained only by false beliefs that one was incapable of correcting because, though there were good reasons against them, one was incapable of responding appropriately to them. To acknowledge them as one's former beliefs or to claim ownership of desires which one approved of on the basis of them may well diminish one's self-respect, or be a source of guilt or shame. "I can't believe I wanted that," may express such a feeling. This may mark an important difference, for insofar as one's desires and the reasons which sustain them are based on reasons, then one can acknowledge them as one's own even after they have been revised or abandoned as a result of discovering better options or acquiring true beliefs. If one could acknowledge them in this sense, then I see no reason to deny that when one did identify with them and approved of their effectiveness, they resulted in acts of self-determination. This, I suspect, marks the difference between choices which are susceptible to change as a result of experience and normal processes of learning and those which are held, not as one's own though intermediate desires, but which are genuinely false and non-autonomous.

(2) Avoidability of Error

Secondly, in determining whether a desire is autonomous or not, even though it is based upon a false belief, we need to know whether the agent could have avoided acquiring the false belief. If an agent could not have avoided forming a belief which is false, then any desires which are approved of on the basis of that belief are not autonomous. Here the focus is on the external context in which beliefs are acquired and desires evaluated, rather than on the psychological component of critical competency which I have called reasons-responsiveness. Any context which makes it inevitable for persons to fall into error undermines the ability of agents to be reasons-responsive, even though it does not impair the actual psychological abilities of the agent; this might be so because, within that context, relevant information is withheld from persons or false beliefs are systematically promoted.

Wolf clearly recognizes that the question of avoidability must be addressed. Thus she writes that what determines whether false beliefs undermine autonomy or not is whether the formation of them was "inevitable given the social circumstances in which they developed".[33] The question that is central to her more recent view is whether or not a person could have formed beliefs in accordance with Reason (where Reason is understood as the highest faculty or set

[33] Wolf 1987, p.146.

of faculties there is, which, in most circumstances, will help us form true beliefs).[34] Only if it is inevitable that people would develop false beliefs should we be persuaded to judge such people as unfree and their relevant desires non-autonomous.

We might wonder, here, whether it is because the desires and beliefs of such people are mistaken or unavoidable that we want to exclude them from full freedom, however. Wolf's answer to this question is that it is unavoidable error that threatens autonomy, that is, that both the falsity of the belief and its inevitability are jointly necessary to defeat autonomy. The reason that the desires and actions of slaveowners, for example, with respect to their interaction with their slaves, are not autonomous is that

> there are certain features of their characters that they cannot avoid even though these features are seriously mistaken, misguided, or bad. This is so because, in our special sense of the term, these characters are less than fully sane. Since these characters lack the ability to know right from wrong, they are unable to revise their characters on the basis of right and wrong, and so their deep selves lack the resources and the reasons that might have served as a basis for self-correction.[35]

Sane selves, on the other hand, though equally determined to have the deep selves they do, are not determined to have seriously mistaken deep selves, for they are capable of self-correction. We are not compelled, by external

---

[34] Wolf 1990, pp. 68-71.

[35] Wolf 1987, p.147. We shall explore Wolf's suggestion that sanity is a necessary condition of autonomy more thoroughly in the next chapter.

circumstances, to keep those characteristics which are objectionable. Thus it is the combination of false beliefs and the inevitability of those beliefs together which exclude such persons from being fully free or autonomous with respect to some of their desires.

Wolf is surely right to insist that desires whose approval is sustained only by unavoidable false beliefs cannot be considered fully autonomous. The avoidability of beliefs marks the difference between socialization and indoctrination, between our society and that of *Brave New World*, and between "normal" and "abnormal" contexts of preference-formation.[36] Contexts of desire-formation which make it inevitable that individuals will acquire false beliefs, when those beliefs lead to the adoption and maintenance of desires, undermine the reasons-responsiveness of agents and hence the autonomy of their desires.

The autonomous person's desires are truly authentic. Such a person can and does subject his desires to critical scrutiny and he alters them when he perceives a balance of reasons in favour of such a revision.[37] Such critical scrutiny requires that autonomous agents be able to examine not just their desires themselves but the reasons (beliefs) which support them. When a person's approval of a desire is based only upon unavoidable false beliefs, on the other

---

[36] The contrast between normal and abnormal contexts of preference-formation is developed in the following chapter.

[37] Joel Feinberg, "Autonomy", in Harm to Self (Oxford University Press, 1986); reprinted in The Inner Citadel ed. John Christman, p.32. Citations of Feinberg 1986 are to the latter collection.

hand, the expressed desires of agents, even though self-identifying, are not autonomous. Such contexts of desire-formation and subsequent evaluation make genuine *critical* scrutiny impossible.

How ought we in general to decide whether individuals have the ability to acquire true beliefs in abnormal contexts, given that they have the ordinary cognitive abilities to acquire such beliefs but operate in a context which sustains false beliefs? I want to suggest that we ought to adopt an objective "reasonable person" standard. Our question, then, ought to be, "Could a reasonable person, with ordinary cognitive abilities, have acquired true beliefs (or corrected his false beliefs) in the social context under discussion?" To know whether a context undermines the capacity of agents to form reasonable beliefs we must examine, not what any particular individual believes or desires, but what beliefs reasonable persons in that context could have formed.[38]

Though I have discussed the possibility of error primarily in connection with the general worry that the context of choice might undermine autonomy, the conclusions drawn here have other implications as well. Any forms of influence which cause those subjected to them to form false beliefs which are unreasonable and unavoidable pose a threat to the autonomy of their desires and so should not be used. (This conclusion rests upon the assumption, unargued for in this work, that autonomy is a value. Even if this assumption is granted, of course, the

---

[38] Stanley I. Benn makes a similar suggestion in "Freedom and Persuasion", Australasian Journal of Philosophy Vol. LXII, No. 161 (July 1967).

conclusion that those forms of influence which undermine autonomy ought not to be used is surely defeasible, and may be over-ridden by other values, for example, by considerations of utility. We would need a substantive moral theory which locates the value of autonomy within a larger normative framework to fill in such details.) Autonomy-inhibiting forms of influence may include those which employ deception, or keeping individuals in ignorance of relevant information. Thus we have reason to endorse Dworkin's intuition that "Methods which rely essentially on deception, or keeping an agent in ignorance of relevant facts, are to be avoided."[39] Presumably both propaganda and censorship would be condemned as "illegitimate" techniques of behaviour-control for this reason. What these considerations point to, furthermore, is that the autonomy of individuals can be interfered with by methods of influence which do not directly undermine their reflective capacities or their ability to form second-order volitions. This indicates that autonomy cannot be analyzed as a wholly internal phenomenon, that autonomy is not just a function of the psychological states of the agent, and that specifying the conditions of procedural independence requires attending not only to the external methods of influencing the desires people come to have and endorse but their beliefs as well.

---

[39] Dworkin 1976, p. 27.

CHAPTER VI

AUTONOMY AND FALSE VALUES

VI.1 INTRODUCTION

I want now to consider the role of false value judgments in the formation of false desires. An agent's beliefs about what is good, what is morally permitted or required, what is valuable and the like often have a great influence on her attitudes toward her desires. Such beliefs must be a central component in "full rationality", as Haworth has described it, for to be fully rational is to take one's desires and ends as objects of evaluation; that evaluation will often draw on the normative judgments of the reflective agent, concerning the worth or value of her desires and their objects. But, on many metaethical positions (realism, relativism and constructivism, for example), such value judgments may be false, and when they are they may lead to the adoption of false desires, often of the second order.

Once again I will contrast the views of internalists (who insist that one's normative judgments must inform one's evaluation of one's ends and desires when they are relevant, but allow that those judgments themselves need not be justified by external standards of correctness) with the position taken by externalists (who insist that one's normative judgments must be correct or justified by external standards if the desires and ends which depend on them are to be autonomous). I will argue that both positions are problematic, and that the

conditions I presented at the end of the last chapter offer better criteria than either internalism or substantive externalism for determining whether a desire is autonomous, where it is approved of because of false normative beliefs which the agent has.

VI.2 INTERNALISM, EXTERNALISM, AND VALUES

The internalist position on values can be seen as informing the work of both Watson and Christman. Watson argues that autonomous action must be motivated by one's values, rather than one's mere appetites or passions. Values, it will be recalled, are those desires which the agent judges to be good, either intrinsically or because the object at which they aim (the state of affairs they would bring about if satisfied) is judged to be good. Together such values constitute one's "evaluative system", and autonomous action is motivated by them. One's values are one's authentic or autonomous reasons for action.

Watson's characterization of values falls within what I have identified as the internalist position, because he imposes no objective truth or reasonableness condition (such as, defensible given one's evidence, though perhaps false) upon the value judgment which distinguishes one's values from one's desires. Thus, he writes, "We might say that an agent's values consist in those principles and ends which he -- in a cool and non-self-deceived moment -- articulates as

definitive of the good, fulfilling, and defensible life."[1] In insisting that one's judgments be made in a cool moment, Watson means only to exclude judgments which are made while in the grip of passion (anger or jealousy, for example) or under the influence of debilitating factors (such as inner compulsions or alcohol). He does not provide any analysis of self-deception, but what he says implies that one must not be deceived about one's own relevant conative and cognitive states. Nonetheless, there is a sense in which the requirement that agents be non-self-deceived seems to rely upon an external standard, for the subjective judgment of the agent surely cannot be the final arbiter on the question of whether he is self-deceived. If the requirement that one be non-self-deceived is taken to imply externalism, though, then all the authors considered in this work are externalists to that extent. We need not embrace this conclusion, however; for the criteria of the truth of one's self-referential beliefs just are one's subjective states. That is, there is no external standard of correctness concerning one's beliefs to appeal to here beyond one's actual psychological states. So even if the agent's judgment concerning his beliefs cannot be taken as authoritative, no external standard is invoked. Externalists insist that the truth of one's beliefs is, at least in part, determined by factors which are independent of one's psychological states; no such external standard is involved in determining whether or not one is self-deceived about one's own cognitive and conative states. I take it, then, that Watson's position is internalist as I have been using that term. Whatever the

---

[1] Watson 1975, p. 116.

(cool, non-self-deceived) individual believes to be good, then, constitutes his values (his autonomous desires, on Watson's account), regardless of whether or not the value judgment is defensible by external standards -- by reference to an objective moral code, the moral code of his community, or even the logical demands of moral reasoning. "One's evaluational system may be said to constitute one's standpoint, the point of view from which one judges the world",[2] but there is no requirement that one's standpoint be reasonable.

Christman, too, defends an internalist position on values.[3] He has stated explicitly that he resists any suggestion that one's value judgments must be defensible by an external standard, adopting instead a "content-neutral" account of autonomous values. On this view, so long as one's values have been adopted in accordance with the procedural conditions of autonomous preference-formation, one is autonomous with respect to them independently of the their content.

Even Haworth adopts an internalist position with respect to values. Accordingly, he resists the suggestion that the critical competence which autonomy demands requires that one's desires conform with externally justified moral norms.

> [W]e are bound to admit the possibility of an extremely autonomous person being highly immoral. Critical competence requires taking a critical attitude toward one's

---

[2] Watson 1975, p. 117.

[3] Cf. Christman 1991, pp. 356-359.

> ends as well as toward the means by which those ends are pursued; but it is not obvious that a condition of being adequately critical is that one choose only moral ends.[4]

I consider this to be an internalist position, because only a subjective standard of the "critical attitude" one must take toward one's ends is compatible with adopting "highly immoral" ends. We shall examine the arguments which support the internalist position shortly, but before that some externalist accounts will be presented.

Campbell, Benson and Wolf will be taken as representatives of the externalist position on values and normative beliefs. Campbell treats false normative judgments in the same way as he treats false factual and instrumental beliefs. If one has a false normative belief, and that belief is the sole reason for an agent's approving of a particular goal or desire, then the corresponding desire or volition is false. As such, it cannot be autonomous. For example, suppose a person were to believe that homosexuality is evil, and that love between partners of the same gender is always offensive from a moral point of view. Such a person might engage in various forms of behaviour as a result of these beliefs: she might shun any gays and lesbians whom she meets, or even engage in "gay bashing". Suppose, further, that she has developed a friendship with another, being ignorant of the fact that her friend is gay. Upon discovering her friend's sexual orientation, she may be expected to desire to end her friendship with him. She will reject her previous feelings of affection for him and her desire for his

---

[4] Haworth 1986, p. 157.

company as reasons for action. Thus she will adopt a second-order volition to resist her previous desires to be friends with such a person. Her volition and new desires are, we shall suppose, based only on her false normative beliefs about the immorality of homosexuality. If she were to develop more correct beliefs, she would have reason to revise her desires. Campbell holds that, in at least some cases, having such false normative beliefs is sufficient to defeat the autonomy of one's volitions and desires (those that have been adopted or maintained just because one has the false belief).

Benson, likewise, defends an externalist conception of autonomous values. He argues that it is a necessary condition of free action and autonomous desires that one have "the ability to criticize courses of action competently by relevant normative standards. This ability lends normative substance to the idea of free action, for it entails that full freedom of action is impossible without a certain appreciation of values."[5] His position differs from Watson's, and ought to be characterized as externalist, because he insists that one's value judgments must be correct. "Contrary to Watson," Benson writes, "I want to argue that freedom is a normative notion to the extent that attributions of free action in any particular context depend for their correctness partly on the *content* of the agent's normative understanding, not just on the agent's having some valuational point of view or

---

[5] Paul Benson, "Freedom and Value", Journal of Philosophy Vol. LXXXIV, No. 9 (1987), p. 469.

other."[6] In a more recent paper, Benson elaborates on his position, arguing that oppressive socialization, in which women develop false normative beliefs concerning the importance of physical appearance to their self-worth, undermines their autonomy to the degree that they have false values and so are unable to competently assess their own worth by correct normative standards.[7]

Finally, Wolf's theory falls in with the externalist point of view. In "Sanity and the Metaphysics of Responsibility", Wolf is concerned that one could satisfy the conditions for autonomy found in the theories of Frankfurt and Watson and yet have seriously mistaken values. Because she believes that, insofar as one's normative beliefs are seriously false or one's values are bad, one cannot be autonomous with respect to them, she argues that the "deep self" view of autonomy offered by Frankfurt and others needs to be supplemented with a condition of (moral) sanity. Using the legal definition of sanity developed in the M'Naughten Rule, to be sane an agent must (1) know what he is doing (the nature of his act) and (2) be capable of knowing that what he is doing is, as the case may be, right or wrong (morally or legally). Wolf thinks that those who desire full freedom, i.e., autonomy as I understand it, must desire that their superficial selves be controlled by their deep selves and that their deep selves be sane. Thus they must desire not just that their nonevaluative beliefs be governed by an accurate perception of the world, but that their evaluative judgments be

---

[6] Benson 1987, p. 472.

[7] Benson 1991.

similarly governed.

> The same goes for the second constituent of sanity - only, in this case, one's hope is that one's values be controlled by processes that afford an accurate conception of the world. Putting these two conditions together, we may understand sanity, then, as the minimally sufficient ability to cognitively and normatively recognize and appreciate the world for what it is.[8]

Wolf argues that certain groups of individuals do not act autonomously -- despite the fact that they act as they truly want to act -- because they have false normative beliefs. Thus she claims that "the slaveowners of the 1850's, the German Nazis of the 1930's, and many male chauvinists of our fathers' generation" may not (in the relevant respects) be autonomous (nor responsible) because they have false beliefs about the moral permissibility of their actions and false values. More recently, Wolf has argued that "normative competence" is necessary for fully free actions. Normative competence "implies the existence of nonarbitrary standards of correctness, standards that are independent of an individual's will and even of an individual's psychology as a whole, by which one can judge some actions, choices, ways of life, or systems of value to be better than others."[9] To be normatively competent is to recognize these standards of normative correctness, and allow them a proper role in one's practical deliberations. To do so is to be governed by "the Right and the Good" in one's choice of ends and motives.

---

[8] Wolf 1987, p.145.

[9] Wolf 1990, p. 124.

This dispute is difficult to adjudicate, for there are arguments in favour of both the internalist and externalist positions.

Christman, for example, argues that there are decisive reasons against imposing an external "value condition" on autonomous choice, of the type argued for by Benson and other substantive externalists. Christman argues against Benson's position on two fronts. First, he raises counter-examples against Benson's view. Consider a confirmed and consistent egoist, who rejects any external values that reflect the interests of others as reasons for action. Presumably one could flesh out the example in such a way that a person could adopt this position without violating any plausible constraints on autonomous preference-formation. Having adopted this normative position, however, the agent could be expected to act in ways which violate the norms that others agree apply to his actions (passing by uncaringly while a child drowns in a wading pool, for example). There seems to be no obvious reason why we must conclude that the egoist's desires and actions are nonautonomous.

Christman's second argument against imposing a condition of objectively correct values upon autonomous desires is grounded in the central sense of autonomy as self-direction and authentic desires as truly one's own. Objectively justified values and norms may be wholly external to the actual motivations of the egoist. To say that acting on the former is the only way to express one's autonomy, while acting on one's own critically adopted but unjustified values cannot be expressive of autonomy, seems, Christman says, "downright

counterintuitive".[10]

Christman mentions further reasons for resisting the suggestion that correct moral values are necessary for autonomy. First, the "correctness" of a person's values is often open to dispute or is indeterminate. Insofar as moral claims are indeterminate, then any account which makes conformity with correct values a necessary condition for freedom or autonomy will render those concepts equally indeterminate.[11] One need not be a complete moral sceptic to see the seriousness of this problem: so long as reasonable people can disagree about a moral question, the ability of people to act freely or autonomously with respect to actions which involve that moral issue cannot be established. As Christman remarks, moral uncertainty typically does not lead to a suspension of judgment concerning the freedom or unfreedom of individuals and their actions. This supports the claim that conformity with correct values is not conceptually necessary for freedom or autonomy:

> There remains a lack of stable consensus on a variety of deep moral questions. However, we don't take that as leaving in doubt the claim that a person in a particular case is acting freely. We don't postpone the question of free action until the moral controversy has subsided. Indeed the fact that the two questions are not even distantly related in our minds is evidence that the meaning of freedom is fixed independently of the determination of correct values. This, in turn, indicates that value commitments are not necessary conditions for freedom.[12]

---

[10] Christman 1991, p. 357.

[11] Christman points to this problem in Christman 1991, p. 358.

[12] Christman 1991, p. 358.

To insist that one must have values and normative beliefs which are "correct" by some objective standard is certainly the most controversial version of the thesis that authentic desires must be supported by true relevant beliefs in order to be autonomous. For it will run afoul of many metaethical theories. Non-cognitivists (of either the emotivist, prescriptivist or subjectivist varieties) would reject the meaningfulness of the claim that a moral belief could be false.

If one takes a cultural-relativist approach to ethics, there are equally serious problems. For if the truth of one's normative judgments is a function of the widely shared convictions of those within one's society, and autonomous desires must conform with those judgments (or, at least, not be contrary to them), then one could never autonomously desire that which is genuinely condemned within one's society. To desire that which is condemned would be evidence both of having made a normative mistake and of being non-autonomous with respect to those desires which are endorsed on the basis of that mistake.

But even if one accepts a realist or cognitivist position, there will be counter-intuitive results. One could never autonomously desire to do wrong: all immoral desires would rest on some kind of normative mistake. To desire that which is immoral would be evidence that one's normative competence is impaired. Immoral actions would be non-autonomous because they must have their source in such an error, or be cases of weakness of will.

Yet there seems no compelling reason to accept any of these claims. Indeed, there are good reasons not to do so. Ignoring for now the problems presented by

non-cognitivist accounts of moral language, it would seem that on either a relativist or realist account of ethics we sometimes allow that people can autonomously act badly. In a society which condemns homosexuality as morally evil, it is possible for a woman to autonomously choose such a sexual orientation. From the vantage point of realism we might be persuaded that certain acts of political terrorism -- random bombings of public places, for example -- are wrong. Yet, a person could well define herself through membership in a political cause which employed such actions, as a form of protest (as does the Irish woman in *The Crying Game*). There seems no good reason to insist that her political desires are nonautonomous, even though the normative beliefs which sustain them (she believes that such actions are morally permissible) are condemned by a strong majority of persons in her own society or by an objective moral code. These possibilities can be represented on the internalist account, but not on the substantive externalist position.

Nevertheless, there are good reasons not to accept wholeheartedly the internalist position on values. Externalists seem to have some powerful arguments on their side of the debate, to which we must now turn. Recalling Campbell's theory of false desires, consider the case of Sara. Sara has never been physically or emotionally attracted to men, and she has not been able to enter into a satisfying relationship with a man. She has been attracted to other women, however, and she identifies herself strongly with her female friends. For these and other reasons she believes that she is a lesbian. And she desires to act on her

homosexual inclinations. Yet Sara has also been raised in a society in which homosexuality is considered perverse and unnatural, and she believes her lesbian desires are evil. Because of her beliefs about the immorality of her desires, she forms a second-order desire to resist them. The conclusion of the internalist bi-level view would be that Sara's desire to resist her lesbian desires is authentic and more truly her own than her lesbian desires themselves, which she has made external to her. Many people would find this conclusion troublesome, I think. But on Campbell's view, this second-order desire is false and so not autonomous, because it is based on and sustained only by a false value judgment. If she came to believe that homosexuality is neither immoral nor unnatural, then this would have a corresponding effect on her second-order desire. Sara's own desires and beliefs would give her good reason to pursue her lesbian desires once her false normative beliefs have been corrected. Here the externalist, rather than the internalist, seems to give the right answer concerning the autonomy of Sara's desires.

Benson has argued, in further support of the externalist position, that autonomous agents must care about whether their actions and reasons for action are justified; this is plausible, and is presupposed by the bi-level theory of autonomy. But, Benson reasons, "If we care deeply about the value of our actions, then we want more than the power to translate our own value judgments into effectual willing. We also want to be able to appreciate the relevant values

and arrive at competent appraisals of the alternate courses of action we face."[13] Thus, autonomous agents must care not just that their desires are supported by reasons and comport with their values, but that their desires and values are supported by good reasons, including correct normative ones when they are relevant.

Furthermore, the externalist position seems to offer a reason why severely deprived upbringing and oppressive socialization inhibit the future freedom and autonomy of those persons who have been subjected to them. If one has acquired seriously false values and normative beliefs because of such influences, the internalist position allows the possibility that those individuals could be autonomous with respect to them nonetheless, provided that they survive the relevant reflective assessment. Both Wolf and Benson deny this possibility, on the grounds that oppressive contexts of socialization deprive people of "the development of the ability to evaluate [their] own conduct competently".[14] Someone who has been prevented from developing normal normative competence does seem to lack a skill which is needed to be fully in control of her own will and actions.

Finally, it is thought that the actions of autonomous agents provide an important kind of self-disclosure: when an autonomous agent acts on an

---

[13] Benson 1987, p. 475.

[14] Benson 1987, p. 478; Wolf 1990, pp. 75-76 and her example of Jo-Jo in Wolf 1987.

autonomous desire, we can draw inferences from her action to the kind of person she is. This seems not to be true in many cases of actions performed by those who are incapable of competent normative evaluation of their options and desires. For example, a child who cheats at a game, while appreciating the relevant norms and rules which govern play and that they apply to her, shows herself to be a cheat; on the other hand, a child who violates the rules, while not appreciating why they matter or apply to her as a player, does not. Thus, it would seem, our common practice of basing evaluations of an autonomous person's character upon her actions presupposes that the person is capable of being governed by a competent appraisal and appreciation of relevant normative facts.[15]

There are, then, reasons in support of both the internalists' and the externalists' positions. A more reasonable view would draw on the strengths of each.

VI.3 A MORE REASONABLE VIEW

The plausibility of the substantive externalist position is more apparent than real, resting as it does on an equivocation (mentioned briefly in the discussion of Wolf's views in the previous chapter) concerning the sense of normative *capacity* which is at issue. Wolf and Benson are surely correct to insist that it is a necessary condition of desire autonomy that a person have the capacity to

---

[15] Cf. Benson 1987.

appreciate the relevant reasons in favour of, as well as against, her various beliefs and desires. To recognize and appreciate the "reasons there are", both evaluative and non-evaluative, is just to be reasons-responsive, as I have characterized that condition. A person who lacks reasons-responsiveness would be insane by many standards, for she would be incapable of knowing that her actions are wrong when they are. Such a person would fail to display autonomy with respect to at least some subset of her desires, for she would lack the capacity to respond to moral reasons, even when they are relevant to the assessment of her desires and values.

But it is deeply problematic to claim that one must actually exercise this normative competence in making one's choices, so that one must choose only in conformity with external standards of correctness to be autonomous. The latter view would imply that if one's practical deliberations were influenced either by false beliefs or false values, or by ignorance of relevant facts which constitute reasons for or against what one proposes to do, then one is not fully free with respect to that action. This latter view is implausible, for it implies that one can never freely (autonomously) choose to do wrong, for to say that an action was wrong is to say that there were sufficient reasons for not doing it. Moreover, one could have reasonable beliefs which fall short of such an external standard of correctness; surely so long as one's beliefs are reasonable, they can ground autonomous desires.

Though I agree with Wolf that sanity is a necessary condition for agent

autonomy, I do not want to rush to embrace the conclusion that all desires which are based on false normative beliefs must be non-autonomous. Only if those beliefs are the inevitable result of defective faculties or one's social environment do I want to say that false values and false evaluative beliefs undermine the autonomy of desires based upon them.

A person may simply have some false desires which are based upon mistaken normative beliefs. If they are moral beliefs which are both false and beyond correction (i.e., not reasons-responsive), then this will undermine the autonomy of the desires which receive support from them; otherwise, they may simply be false yet autonomous. If, as Christman argues, many moral judgments are such that reasonable people could disagree about them, then it ought to be enough that any individual's beliefs about the matter be reasonable, even though at a later time it may be decided that that individual's view was in fact false.

The case is different when a person has unavoidably fallen into moral error, however. This brings to the fore the second of the conditions I listed in the previous chapter as relevant when considering whether false beliefs undermine the autonomy of desires: the avoidability of the error. As we have seen, Wolf also thinks that the question of whether or not one could have avoided making false normative judgments and so could have avoided forming bad values is central. For in cases where one could not have avoided forming bad values, one could not have been governed in one's practical deliberations by the True and the Good. The equivocation between having the capacity to form right judgments and

actually exercising that capacity which, I think, plagues Wolf's writings, however, is evident in her own examples of persons who could not have avoided making moral mistakes and so who are unfree with respect to some of their desires and values.

Her examples are of slaveowners in the 1850's[16], German Nazis in the 1930's, and chauvinists of our father's generation. Now, it seems to me that Wolf's examples do not serve her purpose. Certainly in the case of German Nazis (who had reached maturity before 1933, say), it is not true that they could not acquire true relevant beliefs. There is no reason to suppose that they were "unable to cognitively and normatively recognize and appreciate the world for what it is", though they clearly did not.[17] Must we conclude that these individuals were not autonomous with respect to their racist desires? Wolf answers in the affirmative, because she believes that they were incapable of knowing that their values were bad.

Though I doubt that Wolf's examples serve her ends very well, her general position is, I think, sound. If a social context is such that it makes it inevitable that people will form morally pernicious values, then that context undermines the freedom and autonomy of its members, at least with respect to those values. The issue of the avoidability of moral mistakes is at the heart of Cheshire

---

[16] I shall return to this example shortly.

[17] Wolf 1987, p.146.

Calhoun's distinction between 'normal" and "abnormal moral contexts".[18] When people form false beliefs and engage in wrongdoing in a normal moral context they violate public moral standards. The world makes it possible for them to know and do the right thing, and so if they fail it is due to a personal cognitive or volitional defect. In abnormal moral contexts, by contrast, there are certain social practices and public standards of morality which are themselves morally wrong, and individuals are misguided by those standards. In such contexts, wrong-doing stems from shared moral ignorance and is sustained by its being normal. Here the world encourages participation in wrongdoing. Feminist consciousness requires marking such a distinction, for we need to distinguish between the rapist or pimp, say, who violates social norms and, so, whose moral ignorance is clearly avoidable (given anything less that an abnormally deprived childhood), and those who participate in oppressive social practices whose "harmfulness" is not generally recognized outside of feminist circles. In the latter case, the moral ignorance of the participants is, perhaps, inevitable.

> In normal moral contexts, the rightness or wrongness of different courses of action is "transparent" to individuals, where "transparent" does not mean self-evident, but simply that participants in normal moral contexts share a common moral language, agree for the most part on moral rules, and use similar methods of moral reasoning. ...The sharing of moral knowledge allows us to assume that most rational, reflective people could come to correct judgments about which courses of action would be right, wrong, or

---

[18] Cheshire Calhoun, "Responsibility and Reproach", Ethics 99, (Jan. 1989).

controversial...[19]

Within normal moral contexts we can trust individuals to be self-legislating. This is because the context makes it difficult for individuals to fall into moral ignorance or error in the first place, and even more difficult to sustain that ignorance or error. Moral ignorance in normal contexts is unusual and atypical, and so to explain such ignorance we must be able to make the case that through some personal cognitive defect, due to natural causes or a grossly defective moral education and development, the person could not avoid forming mistaken beliefs. Of course, even if a person has had a normal education, she might remain ignorant concerning specific moral issues. This might be the result of negligence or confused reasoning, for example; but in normal contexts, such ignorance will be neither systemic nor easily sustained over the course of a person's life. A defective moral education, on the other hand, can lead to pervasive ignorance or a thoroughly immoral character.

Abnormal moral contexts differ in many ways from their normal counterparts. Here moral knowledge is neither transparent nor shared. Until ways are found to introduce the new moral knowledge into the shared moral language or the standard of moral reasoning can be revised, moral ignorance and error are virtually unavoidable for most people. Their ignorance, however, cannot be explained by a personal cognitive defect. In such contexts, leaving people to be self-legislating is to perpetuate moral error. Young persons raised in Nazi

[19] Calhoun 1989, pp. 394-395.

Germany, for example, may have been so deformed by the context in which their formative education took place that they could not have avoided falling into pervasive moral error.

The principal problem posed by abnormal moral contexts is not that agents are literally incapable of moral reasoning and taking the moral point of view. They are capable of reassessing the morality of what "everyone else" does.[20] A person who acts wrongly in an abnormal moral context both could and perhaps ought to know better. Yet there are limits to such judgments, for people typically lack any motive to be critically reflective when moral ignorance is the norm. Moreover, moral reflection is needed to examine not only the content of one's beliefs but the modes of moral reasoning which one employs in that reflection as well. Such reflection requires prompting, which will be absent in abnormal moral contexts where there is no common moral language in which the moral knowledge of the minority can become shared. The social acceptance of the practice and the beliefs which sustain it will not only explain the actions of the participants but also excuse them. These facts require that we temper our judgment that such agents "ought to know better". These facts also lead us to acknowledge that in abnormal moral contexts such agents are not fully autonomous with respect to those desires which are based on false beliefs, for here moral error may well be unavoidable.

Thus abnormal moral contexts impose a cognitive defect upon agents from

---

[20] Calhoun 1989, p.398.

without, which rules out the possibility of autonomy at least with respect to a certain subset of their desires and values. This can be brought out by considering the following two cases, which build upon Wolf's claim that slaveowners may have been unfree with respect to their racist values.

Case #1. Consider the case of Mr. White. Mr. White lives in a society in which slavery is widely practised and morally unquestioned. It is the universal belief, let us suppose, among both the slaveowners and their slaves, that slavery is morally acceptable and economically viable as a social institution. Mr. White himself is economically advantaged and owns slaves. He behaves towards his slaves in the manner prescribed by his society. When Mr. White claims ownership of his female slave's child, he does not suffer any conflict within his own motivational scheme; he wants to do so, this desire is not defeated by any other desire and it does not conflict with either his beliefs or his values. Moreover, were he to examine how he came to have this desire and his beliefs he would not find any reason to abandon his desire or to revise his beliefs. By the internal requirements of the bi-level view, Mr. White acts autonomously when he claims ownership of his slave's child. But does he? Many people would say no, although intuitions seem divided, among philosophers at least, on what precisely to say about such cases. What seems clear, though, given the context, is that Mr. White could not reasonably be expected to believe other than he does regarding the institution of slavery and its moral justification. Though, we can suppose, he has the general capacity for normative competence, he could not have avoided

lacking the requisite moral knowledge he needs to correct his false beliefs. So on my account Mr. White did not act autonomously, since he lacked an essential external element in the capacity for that condition.

Case #2. Consider now the case of Mr. Cream. Mr. Cream lives in a society in which slavery has a long institutional history and is himself a slave-owner. Like Mr. White, Mr. Cream has no internal conflict about claiming the child of his female slave as his property and so his action in doing so is autonomous by internal standards. But there is this difference in the cases, and only this difference: Mr. Cream is aware that there is a significant minority of people who question the legitimacy of slavery as a social institution and who challenge the beliefs which sustain it. Mr. Cream's belief in the moral justification of this convention could be an appropriate subject of critical evaluation by him, then, yet he does not engage in such evaluation. On my account of autonomy, Mr. Cream acts autonomously when he claims ownership of his slave's child, even though the beliefs which sustain his desires are false.

Mr. Cream meets all the necessary external conditions for autonomy. We can suppose that he suffers from no debilitating cognitive or volitional defects, and that the context in which he develops his plans and values is such that his mistaken moral views are avoidable. Nonetheless, while he holds to his mistaken normative views, he could defend them as reasonable. Given that his error is wide-spread, he could at least appeal to tradition, social acceptance and the specific cultural beliefs by which the practice of slavery has been defended within

his community; he could, then, offer reasons for his normative views. Furthermore, while the moral knowledge that slavery is repugnant remains confined to a minority of persons, his persistence in his error does not provide evidence against his reasons-responsiveness (indeed, that he can give reasons for his views indicates that he is responsive in the requisite sense). Nothing in the case rules out the possibility that if he came to recognize the objective reasons there are against slavery, he would alter his desires and values accordingly. This is because his error is not the result of a personal defect in his cognitive or volitional faculties, but is caused by his social environment.

No doubt we could complicate our story here by considering degrees to which moral knowledge is accessible in diverse social contexts. How substantial is the minority who have correct moral beliefs? Do they have access to public forums or political power which would assist in the dissemination of their knowledge? Are they a marginalized group? Are the means for dialogue between the minority and the majority in place or must they be forged before a meaningful flow of information can take place? Questions such as these would have to be answered before we could say with any conviction that those in the majority could have and should have corrected their mistaken beliefs. In the case of Mr. Cream, I am assuming that the activities of the minority in his society who condemn slavery were sufficient to have made his moral error avoidable.

I suggested in the previous chapter that, in answering the question of whether or not a mistake was avoidable, we ought to employ a reasonable person

standard. Thus our question ought to be: could Nazis in the 1930's or chauvinists in the 1950's have avoided forming mistaken values and false normative judgments concerning the proper treatment of Jews or women? Though there will be borderline cases, about which reasonable people may disagree, it does not seem that Nazi Germany presents such a case: Nazis could have and should have known better. At least, this seems true of those who had already attained moral maturity in 1933; children raised under the Nazi regime (subjected as they were to extensive propaganda and censorship, and an educational policy which was expressly designed to inculcate in them moral beliefs which were objectively false) could not have been expected to form externally correct moral beliefs. Thus they may be excused for having fallen into moral error, while their parents should not be so excused. Likewise, by this standard Mr. White seems not to have been able to develop more correct views regarding the institution of slavery.

Employing a reasonable person standard may allow us to say (1) that persons in normal contexts could avoid moral error, unless they suffer from such serious defects that they are incapable of attaining the normal level of cognitive and normative competence, (2) that those in abnormal contexts could have avoided or corrected their moral error once a significant challenge had been raised to their erroneous practices, (3) that in a context free of dissenting opinion or challenge, moral error may have been unavoidable even for reasonable persons. Only the latter is a context which rules out the possibility of attaining autonomy, over some subset of their desires and values at least, for those who are socialized in

it.

Recognizing even this limited threat to autonomy from one's social context takes us beyond strictly internalist criteria of autonomous desires. For there is no reason to think that Mr. White's desires would not pass all the internalist tests for autonomy. If we conclude that his desires with respect to his slaves are not autonomous, because they rest on an unavoidable normative mistake, then we have adopted an externalist conception of desire autonomy, though a weaker conception than those we have examined.

CHAPTER VII

AUTONOMY AND FREEDOM OF ACTION

VII.1 INTRODUCTION

In this chapter I shall examine the final point of contention between internalist and externalist theories of autonomy, which concerns the connection between autonomy or freedom of the will, on the one hand, and freedom of action, on the other. Insofar as internalists claim that both agent autonomy and the autonomy of particular desires are wholly determined by the internal psychological states of an individual, they maintain that there is a radical separation between freedom of the will (i.e., the freedom to endorse or reject one's first-order desires as reasons for action) and freedom of action (i.e., the freedom to make one's highest-order desires effective in action). Externalists, by contrast, recognize that limitations on freedom of action, and restrictions of viable options, pose a potential threat to the development of agent autonomy, in extreme cases, and to the autonomy of particular desires more frequently. Thus they reject the thesis that these forms of freedom are radically distinct. I shall argue here that the externalist position is the more plausible of the two.

VII.2 AUTONOMY AND FREEDOM OF ACTION

Frankfurt argues that a person's freedom of the will or autonomy is a function solely of the internal structure of her volitional states, and so I have characterized his view as an internalist one. As we have seen (in Chapter IV), Frankfurt believes that a person is autonomous if and only if she has the capacity to reflect critically upon her own desires and form second-order volitions concerning their effectiveness. A person's desires are autonomous if and only if she wholeheartedly identifies with them as reasons for action. A person's actions are autonomous if and only if they are motivated by her autonomous desires.

Adopting such an internalist conception of the bi-level theory of autonomy, which characterizes autonomy as a function of one's subjective attitudes towards one's desires, leads Frankfurt to recognize only internal threats to one's autonomy. Thus he recognizes that having inconsistent volitions, or suffering from weakness of will, can defeat the autonomy of one's desires and actions. He does not, however, recognize that a person's autonomy can be defeated by wholly external circumstances.

These commitments to the internalist interpretation of autonomy lead Frankfurt to distinguish between freedom of the will -- freedom to endorse or reject one's first-order desires -- and freedom of action -- freedom from external

impediments to doing what one most wants to do.[1] He argues, furthermore, that these forms of freedom are completely independent of one another, that freedom of action is neither necessary nor sufficient for freedom of the will, and so limitations on one's freedom of action in no way threaten one's freedom of the will.

It is clear that Frankfurt is correct in asserting that freedom of action, freedom from external impediments in doing what one wants, is not sufficient for freedom of the will or autonomy. Animals and young children, unwilling drug addicts and kleptomaniacs, may all enjoy complete freedom of action yet be incapable of freedom of the will or of forming autonomous desires. Freedom of action is also clearly insufficient for autonomy as it has been construed in this work, for one could have this kind of freedom without having the capacities which characterize agent autonomy, or while acting on desires which have been rejected after critical evaluation of them.[2] Such considerations have led Frankfurt to write,

> When we ask whether a person's will is free we are not asking whether he is in a position to translate his first-order desires into actions. That is the question of whether he is free to do what he pleases. The question of freedom of the will does not concern the relation between what he does and what he wants

---

[1] This is Frankfurt's gloss on freedom of action, and it is by no means atypical. As we shall see, however, characterizing freedom of action in this way generates certain paradoxical results, and so a more adequate conception will have to be developed.

[2] Cf. Richard Arneson, "Freedom and Desire", Canadian Journal of Philosophy Vol. 15, No. 3 (Sept. 1985); Haworth 1986, Chapter 8.

to do. Rather, it concerns his desires themselves.[3]

Thus Frankfurt maintains that the two concepts of freedom are completely distinct, and that freedom of action is not even necessary for freedom of the will. There is some reason to endorse this further claim. For one may have the capacities needed to structure one's motives through the processes of identifying with some and rejecting others as potential reasons for action, and so be free to form the will one wants, without having the freedom to act on its determinations. That is, one could form autonomous desires without being free to make them effective in action. In such cases external constraints which prevent one's second-order desires from being efficacious do not impugn one's freedom to form such volitions.

Considerable care is needed here, however, for Frankfurt's claim is ambiguous in important ways. If he means only that one can possess the capacities needed for agent autonomy while lacking the freedom of action to carry out all of one's autonomous projects, then he is surely correct. But if he means that agent autonomy (or the capacities necessary for structuring one's motives so as to confer autonomy on some of them) could be developed or sustained even in circumstances in which the freedom of action of agents is severely limited, then he is just as surely wrong. In providing a psychological account of the development of critical competence and the other capacities which are necessary for agent autonomy, Haworth has argued, very persuasively, that the

---

[3] Frankfurt 1971, p. 20.

development of critical competence requires a "domain of autonomy", that is, a domain in which individuals are free to act, to experiment, and to use their self-monitoring and self-critical skills. As those skills expand, so must the domain of autonomy be enlarged. Having open options and some freedom of action is necessary, then, for the development of critical competence and, hence, for agent autonomy.

Freedom of action is necessary for sustaining critical competence as well. In developing normal autonomy, adolescents develop the critical, reflective capacities for examining their own first-order desires, as well as the second-order commitments they have internalized from their parents and significant others. Limited options (or restrictions on the scope of one's freedom of action) pose a serious threat to maintaining those capacities. For, even supposing that one develops the capacities for critically evaluating one's desires and values, and comes to adopt autonomous desires, values, plans and purposes through the exercise of these capacities, if one is seriously constrained in one's options then one will often meet with failure in making one's second-order desires efficacious when they are aimed at objects from which one is restricted. In such cases, the impulse to exercise one's critical competence must be severely retarded. If one has no prospect (now or in the near future) of carrying out one's autonomous projects, of acting on the determinations of one's reflective evaluations, then adopting such motives and engaging in critical reflection will lose their point. Critical reflection upon one's own desires is, after all, a species of practical

reasoning, the conclusion of which is a decision concerning how to act. If one lacks the freedom to act, then such deliberation serves no purpose and so will not likely be sustained. Thus restrictions on one's freedom of action will retard the development and maintenance of normal autonomy.[4]

Haworth is interested not only in the psychological conditions which make autonomy possible, but also in living autonomously. If one concentrates on the conditions of autonomous living, then one will resist any suggestion that autonomy can be analyzed independently of freedom of action. For one lives autonomously, or lives an autonomous life, just in case the plans and purposes, values and desires which characterize that life are truly one's own and such that one approves of them after subjecting them to competent critical evaluation. Haworth's conception of living autonomously goes beyond what I have been calling act autonomy, but it requires that many of one's actions be autonomous as a necessary condition. Clearly freedom of action is necessary for act autonomy, as I have characterized it: to act autonomously one must be free to act on one's autonomous desires. Thus, limitations on the freedom of action of agents, either through physical constraints or the restriction of feasible options, can interfere with autonomy in this sense.

It would seem, then, that freedom of action is necessary for both the development and maintenance of agent autonomy, as well as for act autonomy. Is it also necessary for authenticity or desire autonomy? It would seem so; for

---

[4] Haworth 1986, Chapter 8, esp. 143-144.

restricted options influence desire formation and evaluation in a number of ways. If one believes that one has no chance to pursue a particular goal or is unfree to satisfy some desire which has been endorsed, then this may produce a tension between one's cognitive and volitional states, resulting in dissonance and frustration. One might relieve this frustration by abandoning the desire or devaluing the goal. Of course, whether one feels it necessary to engage in this sort of preference-revision will depend, in part, on whether one believes that one will have the opportunity to act on one's self-identifying desires in the near or even foreseeable future, or whether the opportunity to so act is closed indefinitely or even forever. If one anticipates no opportunity arising which will make it possible to act on one's endorsed desires, then that endorsement serves no effective purpose and so may remain idle or even be withdrawn.

Jon Elster's analysis of the role that restricted options play in the formation and revision of preferences is particularly germane in understanding the threat that restricted options pose to the autonomy of one's desires.[5] He is especially interested in the analysis of the phenomenon of "sour grapes", which he calls "adaptive preference formation" or "adaptive preference change". Preferences shaped by the process of sour grapes he calls "adaptive preferences".

Adaptive preferences have their origin in the need to relieve intrapersonal

---

[5] Jon Elster, "Sour Grapes - Utilitarianism and the Genesis of Wants", Utilitarianism and Beyond eds. A. Sen and B. Williams (Cambridge University Press, 1982); reprinted in The Inner Citadel ed. John Christman. Citations of Elster 1982 are to the latter collection.

dissonance or frustration resulting from the perception that one cannot satisfy one's desires because of external constraints. Consider the following example. A young woman, Lucy, has two suitors who wish to marry her, Bill and Ted. Upon serious reflection, she has decided that she loves and wants to marry Bill. We can suppose that all the internalist conditions of autonomous desire-formation have been satisfied, and so her ranking of the alternatives $A_b > A_t$ is autonomous. But, sadly, it turns out that Bill's birth was improperly registered and so, given the legal traditions of her community, she cannot marry him. Those to whom they appeal for help commiserate with their predicament, but there is nothing that they can do to circumvent the relevant legal provisions. We can expect that Lucy will be distressed at this unfortunate turn of events and, all things considered, she may continue to want to marry Bill. If she comes to believe that the legal authorities are inflexible, however, she may resolve the tension between her desire to marry Bill (which has been endorsed as a reason for action at the second order), and her belief concerning her inability to satisfy that desire, by abandoning the desire, rejecting it as a reason for action, and adopting a new desire, to marry Ted.

There is an intuitive sense here that Lucy is not self-directed, but other-directed (or directed by her external circumstances), in her desires (with respect to her selection of a spouse). Elster's account of adaptive preference-revision provides support for that intuition, in a way that a strictly internalist account of autonomy cannot.

The internalist will have difficulty explaining why adaptive preference-revision results in non-autonomous desires because, after the preference revision has taken place, the new desire may be one which the person not only has but approves of as a reason for action. Thus, after Lucy has revised her preferences, the internalist has no basis for claiming that her desires or subsequent actions on them are non-autonomous. She is unconflicted at the level of her self-identifying desires, and she is acting on desires that she has reflectively adopted and endorsed. Nothing has been hidden from her, and she has not been manipulated in any way. Thus we have no obvious grounds for claiming that the means used to induce this preference-change violate the conditions of procedural independence or include illegitimate external influences. The internalist must conclude, then, that if she now approves of her desire to marry Ted, and rejects her desire for Bill, those desires are autonomous.

Elster would resist this conclusion; for he claims that desires which arise from the mechanism of adaptive preference-formation are not autonomous. Though he provides little direct argument for this conclusion, his analysis provides the basis for such an argument. One reason for thinking that adaptive preferences of this sort are not autonomous is that they tend to be unstable: if the restriction which caused the adaptive preference change were to be removed immediately after the revision was made, then the adaptive preference might be reversed. If the legal barrier in our example were to be lifted, after Lucy had revised her preferences in the way we have described, and so the external impediment to her

pursuing her previous desire to marry Bill was thus removed, then she might reverse her desires. If this were the case, there seems good reason to suppose that what she really wanted to do, all along, was to marry Bill.

Of course, it matters when the restrictions are lifted in our example, for acting on the adaptive desire (though not itself an autonomous action) may come to be autonomous if Lucy discovers new reasons for doing so and approving of doing so, that are not themselves based on the restricted options she faced. It is important to note, though, that as the case has been constructed, no new reasons for or against either her original desire to marry Bill or her adaptive desire to marry Ted have entered into our story. The only thing that has changed is her belief about the options which are available to her. Thus her reversal to her prior preferences is not based on increased experience or on having learned anything new, either about herself or the objects of her desires. Autonomous preference-changes differ from adaptive revisions because, insofar as the former are based on learning new facts or on increased experience, they tend not to be reversible in this way. Thus, if Lucy had come to know Ted better, came to believe that he was in fact better suited to her as a mate than Bill was for reasons she had not considered before, etc., this would give her new reasons for preferring him as a marriage partner over Bill. In this case her preferences are not shaped solely by the perception of external restrictions, and so they could survive the expansion of options she has when the state lifts the legal barrier to her marrying Bill.[6]

---

[6] Elster 1982, Part I.

That adaptive preferences are prone to such dramatic reversals indicates that they are not as fully attributable to a person as others with which she identifies. Furthermore, such reversals would violate the upward monotonic character of autonomous desires.[7] Let me remind the reader that upward monotonicity requires that desires that are autonomous at time t_1, given options {a, b, c}, must still be autonomous at a future time t_2, given options {a, b, c, d, e}, provided that the only difference between t_1 and t_2 is the enlarged set of options. This point can also be made with respect to any particular preference-ordering. Let $\{A_1, A_2 \ldots A_j\}$ be a set of feasible options where $A_1 > A_j$ is an autonomous ordering of some two preferences in this set. Upward monotonicity requires that if the set of options is expanded, the preference ordering $A_1 > A_j$ remains autonomous and the reverse ordering $A_j > A_1$ would not be autonomous, unless new information becomes available regarding these preferences or a new (possibly higher-order) preference emerges, such as the preference to reduce frustration.

The intuition behind the condition of upward monotonicity is that expanded options alone do not disrupt the stability of one's identification with a desire across time. This is not to deny that having expanded options may lead to new experiences and new information which provide new reasons for or against one's previous commitments; nor do I mean to deny that such new experiences can lead to changes in one's autonomous desire set. In saying that the only difference between t_1 and t_2 is the expansion of the set of options open to the person, I mean

---

[7] Cf. the discussion of monotonicity in III.4, pp. 76-79.

that those new options have not yet been explored and so have not led to an autonomous change among her desires based upon new experiences.

This condition of upward monotonicity can be applied to Lucy's reversal of her preferences after the legal restriction on her marriage options has been lifted. At t_1 she had formed the adaptive preferences D_1 (to marry Ted) and $\sim D_2$ (not to marry Bill); or, she preferred $D_1 > D_2$. Her option set, with respect to her marriage partners, included only OD_1 (the option to marry Ted). At t_2 her option set was expanded so as to include OD_1 and OD_2 (the options of marrying either Ted or Bill). This change of options was sufficient to produce a reversal of her desires, so that they now are D_2 and $\sim D_1$; she now prefers $D_2 > D_1$. These changes are ruled out by the upward monotonicity condition.

At the risk of repeating myself, let me remind the reader that I am assuming that no new reasons either for or against D_1 or D_2 have come to light between t_1 and t_2; the only difference between them concerns the option of acting on those desires. This absence of new reasons is vital to understanding the condition of upward monotonicity. If Lucy had (what she took to be) good and sufficient reasons for her desire to marry Bill (and in supposing that this was an autonomous desire, we are supposing that she had such reasons) at t_0, and those reasons have not been outweighed by contrary considerations (as is implied by the assumption that no new experience has been garnered) at t_1, then they are still good reasons for endorsing that desire at t_1. But at t_1 her reasons are defeated by an external limitation on her freedom to act in accordance with them. If that

limitation is lifted, as it is at t_2, then they will again determine her desires. Insofar as her own desires at t_1 have been undermined only by external constraints, as opposed to being overridden by more compelling contrary considerations, her desires at t_1 cannot be considered autonomous at that time.

The instability of adaptive preferences has offered some grounds for saying that desires which are the product of this mechanism are non-autonomous. The real problem with adaptive preferences is not that they are unstable in this way, however, but that the original preference-change (that which is induced by the restriction of options in the first place) is not autonomous. Such preference-changes violate what I shall call the "downward condition of monotonicity". Downward monotonicity (for our purposes) concerns restrictions on the set of feasible options which a person faces. Autonomous desires are downward monotonic.[8] Let $\{A_1, A_2 \ldots A_j\}$ be a set of feasible options where $A_1 > A_j$ is an autonomous ordering of some two preferences in this set. What downward monotonicity requires is this: if A_1 should cease to be feasible, the preference ordering $A_1 > A_j$ remains autonomous and the reverse ordering $A_j > A_1$ would not be autonomous, unless new information becomes available regarding these preferences or a new (possibly higher-order) preference emerges, such as the preference to reduce frustration. Thus an option ceasing to be feasible is not in

---

[8] I am grateful to Richmond Campbell for this formulation of downward monotonicity.

itself a relevant reason for changing one's preferences.[9] The real problem with adaptive preferences is that they violate the condition of downward monotonicity. Mary's original adaptive preference-change between t_0 and t_1 was induced solely by the restriction on her option set, and so the preference-change is not autonomous. The instability of adaptive preferences is properly seen as a consequence of the lack of autonomy at this stage.[10]

The instability of adaptive preferences also provides a test for determining if a desire change is the result of adaptive or autonomy-conferring processes. But as a means for determining if a preference- change has been caused by adaptive or autonomous reasons, the instability of one's desires will not suffice, for often the restriction of one's options will not be lifted and so the opportunity to reverse one's preferences will not surface. Thus, we need a more general test for distinguishing adaptive from autonomous preferences-changes.

Since adaptive preferences are formed just in response to a limitation on a

---

[9] I claimed in III.4 that the account of upward monotonicity re. autonomous desires was analogous to that of validity in classical logic. Validity is obviously not downward monotonic, however, as a reduction in the premise set can destroy the relation of validity between the premises and the conclusion. Consistency, on the other hand, is a downward monotonic relation, since if a set of premises is consistent then every subset of it is also consistent. Consistency is not destroyed by a reduction of the premise set.

[10] It is worth noting that there need be no presumption that the restriction on an individual's set of options itself violated any normative constraints or was objectionable or unjust. No doubt when a person's options are unjustly restricted this gives us further reason to question the autonomy of the person's subsequent desires. But the restriction could come about, as in our example, in an unobjectionable way, and nonetheless undermine the desire autonomy of the individual whose options have been restricted.

person's options, Elster suggests that we need to know what options the person faced, or believed he faced, when forming his desires. This requires, in turn, knowing not only what he was free to do but what he was not free to do:

> We can exclude operationally at least one kind of non-autonomous wants, namely, adaptive preferences, by requiring freedom to do otherwise. If I want to do x, and am free to do x, and free not to do x, then my want cannot be shaped by necessity. (At least this holds for the sense of "being free to do x" in which it implies "knowing that one is free to do x". If this implication is rejected, knowledge of the freedom must be added as an extra premise.) The want may be shaped by all other kinds of disreputable mechanisms, but at least it is not the result of adaptive preference formation. And so we may conclude that, other things being equal, one's freedom is a function of the number and the importance of the things that one (1) wants to do, (2) is free to do and (3) is free not to do.[11]

Elster's proposed definition of freedom combines both internal criteria (concerning wants) and external criteria (concerning options). This is noteworthy, because I have been assuming that we can construe the set of options a person faces in an acceptable way. But there is considerable controversy between those who adopt subjective criteria for counting options (so that something counts as an option only if the person cares about it) and those who adopt objective criteria for counting options (so that something counts as an option if there is some unique action that one can perform corresponding to the option, regardless of whether one wants to perform that action or not). There are serious problems with construing options according to the preferences of persons, however. As

---

[11] Elster 1982, p. 177.

Richard Arneson notes,

> If freedom is to be measured by counting a person's options as weighted by their importance to that person, then quite obviously one possible strategy for enlarging one's freedom is to bring it about that one takes an enlarged view of the significance of one's available options. In Berlin's words, "If I find that I am able to do little or nothing of what I wish, I need only contract or extinguish my wishes, and I am made free." This stoic strategy Berlin labels the "retreat to an inner citadel", and characterizes it a "sublime" form of the "doctrine of sour grapes"; for him any analysis of the concept of freedom that permits freedom to be increased in this way thereby shows its inadequacy.[12]

Berlin's contrast between two slaves,[13] one contented and the other not, vividly illustrates the problem with measuring freedom according to the options one is free to pursue and cares about. For the contented slave adapts his preferences so that they correspond precisely to those of his master and range only over the options which are in fact open to him. The discontented slave, on the other hand, has many desires he is not free to pursue. If freedom simply measures the ability of a person to pursue the desires he cares about, then the happy slave is fully free while the discontented slave is not. Again, this seems highly counter-intuitive.[14] Furthermore, if one's options are simply a function of one's subjective preferences, then changing one's preferences (one's weighting, in Arneson's terminology) literally changes one's options. That one can change

---

[12] Arneson 1985, p. 428.

[13] See for more details II.2 and III.4.

[14] Berlin 1969, pp. 139-140; see also Arneson 1985, p. 428.

one's options simply by changing one's psychological states seems implausible; anyone who would claim this must surely rely upon a stipulative definition of options which deviates widely from everyday usage.

There would be equally serious problems with a wholly objective method of construing options independently of what the person cares about. If we say that one's freedom is a function of the number of actions he is free to perform,[15] or the number of states of affairs he can bring about, regardless of whether he wants to perform those actions or bring those states about, equally counter-intuitive implications are licensed. A person, *A*, could be free to do infinitely many things, even if he wants to do none of them; the set of actions he is not free to perform may include just those actions he wants to perform. Such a person would be just as free as another, *B*, who is free to do infinitely many things, including many of the things he most cares about. From an external perspective, both *A* and *B* would be freer than another person, *C*, who enjoys only a small number of open options, i.e., is free to perform only a small number of actions, but whose options make possible the pursuit of all his self-identifying projects. The conclusions that *A* and *B* are equally free, and that they are freer than *C*, are surely counter-intuitive. The mere fact that one is free to perform some action, even if one does not desire to perform it, or perhaps even believes that doing so would be immoral, seems not to contribute to one's freedom in any interesting way. (Of

---

[15] This is problematic itself, because it raises questions concerning how to individuate actions.

course, one may value having the open option to act immorally, perhaps because one believes withstanding temptation is necessary to maintain a good moral character. Or one may value having open options that one does not now want to exercise because one recognizes that one's desires may change in the future. In such cases, though, it is because of one's current desires that having more options than one plans or desires to exercise is valuable.)

Elster's characterization of freedom, as being "free to do all those things that one autonomously wants to do"[16], as well as his three-part formulation offered just above, includes both objective criteria and subjective criteria (open options and desires which are not themselves the product of other, closed options). Thus he adopts what I shall call the Autonomous Desire Thesis: The amount of freedom a person possesses varies directly with the number and the importance of the things he autonomously wants to do, together with the extent to which his autonomous desires (or personal values) are satisfiable under the options available to him.[17]

Adopting the autonomous desire thesis would provide us with a method for individuating options. It would also show that there is a (surprising) connection between freedom of action and autonomy, i.e., that agent and desire autonomy are necessary for freedom of action. For if freedom of action is the freedom to

---

[16] Elster 1982, p. 177.

[17] Contrast this formulation with the "desire thesis" described by Arneson. Arneson 1985, p. 431. My view is similar to that offered by Haworth. Cf. Haworth 1986, Chapters 7-9.

satisfy one's autonomous desires, then one must be able to form autonomous desires in order to enjoy such freedom. Thus our initial endorsement of Frankfurt's claim, that one can enjoy freedom of action without being autonomous, needs revision. The problem with Frankfurt's account is that he adopts a purely subjective method for individuating options or measuring the scope of one's freedom of action: freedom of action, remember, is just the freedom to do what one wants. Because this formulation of freedom is inadequate, generating as it does all of the happy slave paradoxes, it must be abandoned in favour of something like Elster's account. Once we adopt a more adequate method for characterizing freedom of action, however, it becomes clear that only autonomous agents can enjoy that freedom. Perhaps the freedom that young children and kleptomaniacs enjoy should be characterized as "freedom of movement" rather than freedom of action.

I suspect that Frankfurt has overstated even his own position when he asserts that freedom of action is not necessary for freedom of the will.[18] For to have a free will is to have the *effective* desires one wants, to approve of the desires that actually lead one to action. For Frankfurt, freedom of the will consists in having the will one wants, that is, in coherence between one's effective desires and volitions. Thus Frankfurt seems to be presupposing that agents with free will also enjoy freedom of action, rather than demonstrating the latter's irrelevance to questions of freedom of the will.

---

[18] Frankfurt 1971, p. 20.

It seems sensible to presuppose freedom of action (in Frankfurt's sense) in our discussions of autonomy and freedom of the will, moreover; for without the presupposition that one is free to act on at least some of the desires one endorses, freedom of the will seems not to be a kind of freedom worth wanting (as Dan Dennett might say).[19] It would be cold comfort to tell a prisoner or slave that, despite being unfree to act on his desires, he nonetheless has a valuable form of freedom in being able to structure his desires in such a way as to make some of them authentic.

The kind of freedom of action that is enjoyed by autonomous agents seems to be a kind of freedom worth wanting. Surely the corresponding sense of act autonomy captures the intuitive sense of being self-directed in what one does, moreover, in a way that the mere freedom to form second-order volitions in situations in which one's options are so constrained as to make them inefficacious does not. To say that one is autonomous just in virtue of being able to form second-order desires, even though there are external constraints in place which will prevent one from acting on those motives, seems to imply that autonomy is a wholly internal phenomenon, radically disconnected from one's actual situation in which an autonomous life must be lived.

Freedom of action is not sufficient for act autonomy, however, even if we suppose that the person who enjoys such freedom is an autonomous agent who

---

[19] See Daniel Dennett, Elbow Room: The Varieties of Free Will Worth Wanting (MIT Press, 1984).

has developed autonomous projects, for one's ability to act on the determination of one's second-order volitions may be defeated by internal conditions as well as by external constraints. Instances of weakness of will provide paradigms of internal obstacles to the performance of autonomous actions.[20]

VII.3 COERCION:

There is a more serious weakness with the internalist position as well. Insofar as the internalist does not recognize that wholly external constraints upon a person's options can undermine the autonomy of his desires, he will have difficulty explaining how coercion undermines autonomy. I shall explore this difficulty further here, arguing that an externalist position is needed to explain the threat that coercion poses to autonomy.

Gerald Dworkin takes coercion to be a paradigm of the kind of influence which undermines autonomy, and he is surely not alone in this judgment. Despite this consensus, however, coercion is notoriously difficult to analyze in general, and it is especially difficult to explain how coercion undermines

---

[20] The range of internal defects which undermine one's capacity to carry out one's autonomous projects is very interesting, and has received considerable attention in the literature. They are treated with particular sensitivity in Campbell 1979. Haworth argues that "self-control" is a necessary condition for achieving normal autonomy, because one must able to resist such internal impulses when they would deflect one from one's autonomous projects. Cf. Haworth 1986, Part I.

autonomy on the internalist version of the bi-level view.[21] It is the latter problem that I want to explore here.

I shall begin by adopting Wright Neely's account of coercion:

> Coercion is a matter of one agent's deliberately altering the external circumstances of another in such a way as to render one (or more) of the latter's relatively low-priority desires incompatible with one (or more) of his high-priority desires for the purpose of getting him to do something which, since it violates those low-priority desires, the coerced agent would not otherwise do.[22]

Thus, in a mugging the victim is presented with a choice between retaining his life or retaining his wallet. In the pre-coercion state, the victim wanted to retain both his life and his money, and could jointly satisfy these desires. In the post-coercion state, he is no longer able to satisfy at least one desire that he could have satisfied had he not been coerced.[23] To employ Benn's terminology, coercion attaches new opportunity costs to at least one option which it would not have had but for the intervention of the coercer[24]; thus, the mugger makes giving up his wallet the opportunity cost of his victim's retaining his life.

Actually, this account of coercion needs to be qualified in one important

---

[21] For a good discussion of the general problem involved in analyzing coerced action as unfree, see Thalberg 1978, as well as Gerald Dworkin, "Acting Freely", Nous Vol. 4 (1970). I am indebted to both of these essays in the following section, as well as to Campbell's discussion of coercion in Campbell 1979.

[22] Neely 1974, p. 50.

[23] Campbell 1979, pp. 166-170.

[24] Cf. Benn 1976.

respect. It is not strictly necessary that the coercer in fact alter the external circumstances of his victim in the ways described, but that his victim believe that they have been so changed. As Neely says, "coercion is not a matter of actually altering the coerced agent's external circumstances but rather a matter of altering the agent's *beliefs* about those circumstances."[25] Thus one can be coerced by a highwayman toting a toy gun as well as by a real one.

We might think of this analysis of coercion as an externalist analysis, because it concentrates on features of the victim's external circumstances, rather than on the psychological states of the victim alone. But this may seem at odds with the point just made, namely, that coercion involves manipulating the beliefs of an agent rather than actually changing her circumstances. Notice, though, that we would be inclined to consider a case one of coercion only if the intervention of another induced the belief that one could no longer satisfy at least one desire that one could previously have satisfied. If a natural disaster made it impossible for a person to satisfy some desire which she could have satisfied but for that event, we do not say that the natural disaster coerced the person. Consider, too, that a person could come to believe that another had imposed a new opportunity cost upon the satisfaction of some desire, but unless that belief has actually been caused by a coercive threat (or offer) by the other, we will judge the person to be suffering from paranoia or some other delusion; we will not see this as a genuine case of coercion. Thus the intervention of another is necessary, and this

---

[25] Neely 1974, p. 50.

intervention is a factor external to the psychological states of the victim.

This analysis of coercion has obvious merits. But if we adopt it, then equally obvious problems will arise in accounting for the widely held intuition that coerced actions are not autonomous. For once the victim is in the coercive situation, he will want to do what his coercer wants him to do. This problem has led Frankfurt to reject such external analyses of coercion, and to adopt instead a view which reduces coercion to compulsion. His writings are instructive, for they show the lengths to which one must go to account for the nonautonomy of coerced actions if one analyzes autonomy from a wholly internal perspective.

Frankfurt claims that coercion only defeats autonomy if the coercive threat or offer results in an irresistible desire in the agent to do what the coercer wants and that that desire is one which the agent has a higher-order desire not to be effective:

> Coercing someone into performing a certain action cannot be ... merely a matter of getting him to perform the action by means of a threat. A person who is coerced is *compelled* to do what he does. He has *no choice* but to do it.[26]

Thus on Frankfurt's analysis, to establish that one has been coerced it is not enough to show that one has "no reasonable choice" other than to comply; nor is

---

[26] Frankfurt, "Coercion and Moral Responsibility", Essays on Freedom of Action ed. Ted Honderich (Routledge & Kegan Paul, 1973); reprinted in The Importance of What We Care About, p. 36. Citations of Frankfurt 1973 are to the latter collection. Frankfurt is not alone in arguing that to be coerced it is necessary that one act under a compulsion which leaves one no choice but to comply. Such a position is also argued for by A. Grunbaum, "Free Will and the Laws of Human Nature", American Philosophical Quarterly Vol. VIII (1971), pp. 303-304.

it enough that a threat gives rise to an irresistible desire to comply. The irresistibility of the desire to comply with a threat is a necessary but not sufficient to make a threat coercive. Frankfurt argues that what more is needed, to make a threat or offer coercive, is that the person who submits to his desire should resent doing so, should wish not to be motivated by the desire that moves him:

> An offer is coercive ... when the person who receives it is moved into compliance by a desire which is irresistible but which he would overcome if he could. In that case the desire which drives the person is *a* desire by which he does not want to be driven. When he loses the conflict with himself, the result is that he is motivated against his own will to do what he does.[27]

Thus, for Frankfurt, one is coerced only if one could not have done other than one does and one would not have done what one does except for the coercive offer or threat. Furthermore, one must resent acting on the irresistible desire to comply.[28]

But why must we conclude that a person who submits to coercion resents acting in the way he does? In the case of coercive threats, Frankfurt's answer is that one submits just in order to avoid the threatened penalty. "That is, his [the victim's] motive is not to improve his condition but to keep it from becoming

---

[27] Frankfurt 1973, pp. 41-42; see also Harry Frankfurt, "Three Concepts of Free Action", Proceedings of the Aristotelian Society (1975); reprinted in The Importance of What We Care About. Citations of Frankfurt 1975 are from the latter collection. Dworkin also argues that if an action is coerced it must be motivated by a desire that the victim wants not to act on. Cf. Dworkin 1970, p. 372.

[28] Frankfurt 1969, p. 10.

worse. This seems sufficient to account for the fact that he would prefer to have a different motive for acting."[29] Hence, acting solely to avoid a penalty must be a motive which persons resent acting on.

It is important to keep in mind that the irresistibility of the desire to comply and the resentment the agent feels toward acting on that desire are separate conditions, each of which Frankfurt claims is necessary. The condition that one must resent one's effective motive does not itself imply that the agent is literally overwhelmed by it. One could decide to comply, after careful and calm reflection upon the alternatives, for purely prudential reasons. That the coercer has interfered with the victim in such a way that the victim must now act just to avoid a penalty, or to keep his condition from being worse, could generate resentment independently of the strength of the desire to comply.

There are a number of problems with Frankfurt's account of coercion. Both his treatment of resentment and his insistence that the desire to comply must be irresistible if the compliance is to count as coerced are highly unintuitive. That the desire to comply with a coercive threat must be irresistible for one's compliance to be countenanced as coerced seems to fly in the face of our intuitions concerning paradigmatic cases of coercion. The traveller facing the highwayman, or the bank teller facing the armed robber, or the politician who is being blackmailed, or the diplomat whose family is being threatened, all seem to be cases of genuine coercion. Yet there is no reason to think that they are literally

---

[29] Frankfurt 1973, p. 44.

overwhelmed by the desire to avoid the threatened penalty should they refuse to comply with the wishes of the person making the threat. Deliberation need not come to a halt when one is placed in such situations; there are alternatives to compliance, even if they would be rejected by most people in such situations. One might be literally paralysed by the threat such coercers make, but one need not be so debilitated to recognize the threats as coercive. Yet Frankfurt insists that unless the desire to comply is irresistible, literally compulsive, one's compliance has not been coerced.

In all these cases, the coercer has made it impossible for the victim to satisfy each of the desires which could have been jointly satisfied in the pre-threat situation (keeping one's life and one's money or the bank's money, maintaining one's public office and reputation while also keeping one's money or awarding contracts on the basis of merit or whatever, keeping state secrets to which one is privy and maintaining the safety of one's loved ones). If the threat is credible and the harm threatened believed to be serious, the decision to sacrifice the lower-priority desire in order to satisfy the higher-priority desire might be easily reached, without being irresistible. Frankfurt cannot admit these possibilities because he is committed to the view that a person's autonomy or free will can only be defeated from within -- that questions of freedom of the will and freedom of action are radically distinct, and that whether one is autonomous or not depends just on the internal structure of one's desires. This forces him to identify coercion with compulsion, for he must move from an external threat to an

internal compulsion if he is to account for coercion's interference with autonomy. On his account, the external threat creates an overwhelming desire to comply, but it is only because this is a desire that one wants not to be effective and resents acting on that it defeats one's autonomy.

There are difficulties, too, with Frankfurt's insistence that the victim of coercion must have a second-order desire not to be moved by the desire inspired by the coercive threat. This is just counter-intuitive. The bank teller, for example, who hands over the money to an armed robber, need have no second-order desire to resist her first-order desire to do so, i.e., to comply. One might suggest that she may want to resist what she perceives as cowardly motives on her part, or she may want to resist to be a hero, but surely she need not (either at the time of acting or latter) feel any such higher-order desires to resist her effective desire to comply. She may reasonably believe that her job or employer are not worth risking grave personal injury for, or that compliance is the only reasonable course of action. In such cases, it is difficult to understand why she must have a second-order desire to resist her compliant first-order desire to give the robber the money and save her own life. Such considerations led Irving Thalberg to argue that resort to a split-level analysis of the will distorts our understanding of what it is to act under coercion. He argues that one need not suppose, as (internalist) bi-level accounts must if coerced actions are not autonomous, that the agent who complies with a coercive threat has a desire, at the time of compliance or afterward, to resist acting on the desire which actually motivates compliance. The

desire which motivates action in the most extreme cases is thought to be something like a desire for self-preservation. And it is simply not true that agents generally have second-order desires not to act on such a motive.[30] Nor is it generally true that agents resent having or acting from the desire to save their own lives.

Dworkin seems to recognize that coercion motivates compliance through such basic drives as the desire for self-preservation. Thus, he writes that "Coercion always involves ... utilizing basic drives which almost everyone shares -- self-preservation, avoidance of pain, embarrassment, concern for the welfare of those close to us."[31] Coercion operates by making the satisfaction of such basic, pre-existing drives dependent upon acting in violation of a lower-priority desire, which the coercer has made an opportunity cost of satisfying the higher-priority drive.

If coercion utilizes basic drives in this way, though, how are we to explain the sense in which the coerced agent does not want to do what he is coerced into doing? Frankfurt and Dworkin offer similar answers: in complying with a coercive threat, the victim is acting not to improve his condition, but to keep it from becoming worse. When one acts to achieve some good for oneself, or because one feels required to do so because of some principle one accepts, then we might expect one to approve of one's reasons for action. This is not true, says

---

[30] Thalberg 1978.

[31] Dworkin 1970, p. 375.

Dworkin, when one acts because of fear in situations of coercion:

> Men resent acting for certain reasons; they would not choose to be motivated in certain ways. They mind acting simply in order to preserve a present level of welfare against diminution by another. They resent acting simply in order to avoid unpleasant consequences with no attendant promotion of their own interests and welfare.[32]

This is not a very promising answer. People often willingly engage in activities which are designed only to prevent their present level of welfare from falling (exercising and adopting proper eating habits, to prevent a deterioration in health, for example). A person who moves out of the path of an oncoming truck has been motivated just by a desire to preserve her present level of welfare against diminution by the truck-driver, but (assuming she does not think that he is deliberately or negligently endangering her welfare) she need not resent either having or acting from that motive. By contrast, a person may resent very much being made to do something which does promote her welfare, as in cases of paternalistic interference by others in her life. What one resents, in cases of coercion, is not being motivated by a desire to preserve one's well-being or satisfy the basic drives that Dworkin identifies, but the actions of one's coercer which have deliberately put these things at risk or attached new opportunity costs to satisfying them. What is important in cases of coercion is that someone else has deliberately manipulated one's external circumstances in such a way that one cannot enjoy the same level of welfare that one could have enjoyed without the

---

[32] Dworkin 1970, p. 377.

coercer's interference. Even if one wants to comply with the coercive offer in the post-threat situation, then, one may resent the coercer's interference for this reason.

Resentment can be accounted for in another way as well, which does not require that the victim have a defeated second-order desire not to act on the desire to comply. For coercion involves, among other things, the manipulation of the choice-situation in such a way that no matter what the victim now chooses he will be worse off than he would have been without this manipulation.[33] This is so because the satisfaction of some desire which one approves of is now excluded from one's feasible set of options. One might, prudently, decide to comply with a coercer's demands, but in acquiescing one must sacrifice at least one desire or plan that one wanted in the pre-threat situation. Resentment is the natural emotional response to another's intentionally restricting one's options in a way that interferes with one's ability to satisfy one's autonomous desires.[34]

The discussion of adaptive preference-revision in the preceding section offers

---

[33] This evaluation must be made relative to the *status quo ante*, of course, and there are serious difficulties in determining what the baseline for comparison ought to be between the pre-and-post-threat situation. Cf. Robert Nozick, "Coercion", Philosophy, Science and Method: Essays in Honor of Ernest Nagel eds. S. Morgenbesser, P. Suppes and M. White (St. Martin's Press, 1969). Frankfurt denies that such comparisons are morally significant in Frankfurt 1973, because all that matters on his view is whether one acts freely in the post-threat situation, and that is determined just by one's attitudes toward one's reasons for action at that time. Cf. Campbell 1979 for a criticism of Frankfurt's view here.

[34] Cf. Peter Strawson, "Freedom and Resentment", Proceedings of the British Academy Vol. XLVIII (1960).

a different sort of reason for thinking that coercion interferes with autonomy, even if the desire to comply is not irresistible and the victim has no second-order volition to resist it. In the pre-threat situation a person could satisfy at least two desires (which, we shall suppose, are themselves autonomous); in the post-threat situation, the person can satisfy at most one of the two. Thus the coercer has restricted the victim's set of feasible options. Now we might represent the victim's desire to comply in the post-threat situation (the desire to hand over her money to the highwayman, say) as an adaptive preference. This new desire has been formed only by her recognition that her options have been restricted: it would not have been formed but for that restriction, and it is reversible in the way that adaptive preferences are. If she came to believe that she could, in fact, satisfy her pre-threat desires to keep her money and maintain her physical security, the desire to retain the money would be revived. Her desire to turn over the money seems, then, to have all the marks of an adaptive preference.[35] If adaptive preferences are nonautonomous, as we concluded in the preceding section, then the desire to turn over the money to the highwayman is equally nonautonomous. Thus we have a direct argument, which does not appeal to the

---

[35] This is true as I have characterized adaptive preferences, but not as Elster does. For he claims that the causal mechanism which produces adaptive preferences operates unconsciously. He believes that this additional condition is needed to distinguish adaptive preference-revision from deliberate character-planning, in which one is consciously motivated to revise one's preference set in the face of limited options so as to allow one to satisfy a greater number of one's desires. Cf. Elster 1982, p. 174. I think that this distinction can be drawn without this controversial condition of unconscious causal mechanisms operating "behind one's back". I will say more about this later in this section.

resentment of the agent toward her own motivations, for saying that desires caused by coercive threats are not autonomous.

I have offered arguments in the last two sections for thinking that adaptive preferences are not autonomous desires and that desires which are reversed just in response to restrictions of one's feasible option set thereby undermine desire autonomy. Before concluding this chapter, a potential objection to this view must be addressed. It might be inferred from the claim that desires which are formed in response to limited options are nonautonomous that autonomy requires a kind of oblivion to the real options that one faces, for if one revises one's desires in light of the options one faces, they are said to be adaptive desires and so nonautonomous. Such an inference, if warranted, would show that the account of adaptive preferences developed here is implausible. For one's options are relevant facts to which autonomous agents must, surely, be sensitive in adopting autonomous plans and projects. There seems to be a tension, then, between the demand that autonomous agents be reasons-responsive and the claim that desires resulting from adaptive preference-formation are non-autonomous. Someone who formed her desires without attending to the options which are open to her, or without considering the external constraints which will limit her ability to successfully execute her plans, would fail to be responsive to reasons which are relevant to the assessment of those various plans, yet such insensitivity seems to be required by the account of adaptive preference-revision and monotonicity I have given.

This objection can be defused by considering that adaptive preferences are formed *only* because one faces restricted options. One can engage in what Elster calls "deliberate character planning", in which one decides that one wants to revise one's preference set so as to allow one to satisfy more of one's desires. One might adopt a second-order desire to desire only those things that one has a reasonable chance of acquiring. But such a decision cannot be motivated solely by the perception that one faces limited options. Other considerations must also play a role: that one dislikes being frustrated, that one considers being satisfied more important than the pursuit of goals which one might never attain, etc.. Someone may have such commitments, wanting contentment as a basic goal, and so such deliberate character planning would be rational for such a person. But such planning is supported by reasons which go beyond the belief that one's options have been restricted, and it is aimed at an increase in one's overall well-being. Adaptive preference-revision is not justified by these additional reasons, and the reversibility of the process indicates that one does not take the adaptive preference-change to be such that it will produce an over-all increase in one's welfare.

Similarly, we can see that the monotonicity condition does not run afoul of the requirement that agents be reasons-responsive. An increase in one's feasible options may make the adoption of new autonomous desires possible, and pursuing new goals may lead one to abandon those which were formerly approved of. Such revisions of one's autonomous desires will be based upon the

discovery of reasons in favour of pursuing these new goals or plans, and these may supplant one's previous commitments. All that the upward monotonicity conditions rules out is preference changes which are based just on an expansion of one's set of options.

Externalists deny that the concept of autonomy can be analyzed just in terms of the subjective attitudes of agents toward their own desires. They recognize that external influences can undermine a person's autonomy without thereby undermining the subjective conditions of autonomy (such as the ability to reflect upon her desires). They do not draw any radical separation between freedom of the will and freedom of action, therefore. This allows externalists to recognize that restrictions on agents' freedom of action can interfere not only with their ability to act autonomously, but to form autonomous desires as well. Adopting the externalist position also provides grounds for claiming that persons are not motivated by autonomous desires when they comply with coercive threats, even though there is a clear sense in which they want to comply so as to avoid the threatened penalty.

CHAPTER VIII

CONCLUSION

VIII.1 INTRODUCTION

The principal aim of this thesis has been to provide a philosophical analysis of personal autonomy. As a piece of conceptual analysis, it is rather old-fashioned in its approach to philosophical questions. Here I want to present a brief summary of the theory of autonomy which has emerged in the preceding chapters. I shall then argue that this theory can meet the central objections which have been raised to both the internalist and substantive externalist theories which were examined.

VIII.2 A SUMMARY OF THE THEORY

Personal autonomy is the condition of being self-directed. Self-direction, I argued, requires that one's actions be motivated by authentic desires. Autonomous action is motivated by authentic reasons for action.

The attainment of autonomy requires, then, the development of authentic reasons for action. Authenticity was explicated by utilizing a bi-level theory of desires: one makes some of one's first-order desires, projects and values authentic by critically reflecting upon and identifying with them as reasons for action. One

identifies with a particular desire when one approves of it as a reason for action, thus desiring that it be an effective desire (i.e., making it the object of a second-order volition). Provided that one's identification with a desire is decisive and one is not ambivalent at the second order with respect to it, this sort of second-order endorsement of a desire is sufficient to confer authenticity upon it. This account of identification provides the tools for distinguishing authentic desires from those which one has acquired uncritically from others, pre-reflective (or non-reflective) passions and impulses, and desires whose occurrence one experiences but that one wishes not to actually lead to action.

The capacity to adopt second-order volitions, to identify with and reject some desires as reasons for action, is necessary for authenticity. And authenticity, I argued, is necessary for autonomy. But authenticity is not sufficient for autonomy. One could have authentic desires but fail to be autonomous with respect to them for a variety of reasons: because one's identification was based on ignorance of relevant facts, or unavoidable false beliefs or false values, or the result of manipulation, coercion, or restricted options. Such factors, I argued, can result in non-autonomous desires, despite the fact that one has the psychological capacities needed for adopting authentic desires and has exercised them in actually identifying with some desires and rejecting others.

Furthermore, the adoption of authentic desires does not ensure that one's actions will be autonomous, i.e., that one will succeed in actually being self-directed. For one can fail to make those desires with which one identifies at the

second-order effective in determining one's actions. This may be due to internal volitional deficiencies, such as ambivalence and weakness of will, or the result of external barriers to the pursuit of one's authentic desires. Thus authenticity is insufficient to ensure the autonomy of either one's desires or actions.

These observations draw upon the distinction I made between agent autonomy, desire autonomy and act autonomy. Agent autonomy is autonomy as it is predicated of agents. It consists of the psychological (cognitive, conative and affective) capacities needed for critical self-evaluation and self-definition. This cluster of capacities I called "critical competence". The critical competence needed for agent autonomy requires authenticity. But authenticity (being a function solely of the psychological states of an agent) is not sufficient for agent autonomy. It is also necessary that the agent be rational and reasons-responsive. That is, an agent must have the capacities to respond to whatever facts (about herself, the world or morality) are relevant to the assessment of her desires, plans and values. It is true that reasons-responsiveness is itself a psychological condition, but it requires that agents be capable of responding to relevant facts, whose truth is itself independent of her psychological states and subjective attitudes. These conditions are taken to be sufficient for agent autonomy, i.e., sufficient to determine that one who satisfies them is an autonomous agent, capable of self-direction.

Yet autonomy as it is predicated of agents cannot be predicated of their desires or actions unless further conditions are satisfied. It is necessary that one

be an autonomous agent to have either autonomous desires or to act autonomously, but it is not sufficient. For a desire to be autonomous the following conditions must be satisfied:

1) it must be the desire of an autonomous agent;
2) the agent must approve of it at the second order as a reason for action;
3) the agent must not be ambivalent at the second order concerning it;
4) the agent must be reasons-responsive with respect to it;
5) the agent's approval of it must satisfy the conditions of monotonicity;
6) the agent's approval of it must not be based on unavoidable cognitive or normative error or ignorance;
7) the agent's approval of it must not be induced by coercion.

Each of these conditions was defended in the course of this work. Together, they are sufficient for the autonomy of one's desires. An autonomous action is one that is motivated by a desire which meets these conditions.

This account of autonomous desires recognizes as necessary certain psychological conditions (the reflective capacities needed to adopt authentic desires, lack of ambivalence with respect to a desire at the second order, and the capacity to respond appropriately to facts which are relevant to the assessment of one's desires and values); but it also requires conditions that are independent, at least in part, of the psychological states of the agent (the conditions of monotonicity and freedom from coercion necessarily make reference to her external circumstances, for example). Because the autonomy of one's desires is a function, at least in part, of factors which are external to one's psychological states, I have called my theory an externalist one. Externalist theories were contrasted with those which were called internalist, because the latter make autonomy depend solely upon the internal psychological states of the agent.

VIII.3 SOME OBJECTIONS RECONSIDERED

It remains to be seen whether my account can meet the objections which I raised against both internalist and other externalist theories of autonomy. Let me begin to answer this question by re-examining Dworkin's worries about his original theory of autonomy (II.2). The first concern that Dworkin raised, and the central reason I offered for rejecting any account which makes the authenticity of a desire sufficient for its autonomy, is that one can be heteronomous at the level of one's second-order volitions. That is, one could decisively and unambivalently approve of a desire as reason for action (and hence it would be authentic) because one has been manipulated or deceived by others. Could one also have autonomous desires, as I have characterized them, which are the product of manipulation or deception? I think not.

Many forms of manipulation or deception are successful as means of behaviour control only because they result in false beliefs being held by the person who is manipulated. This is obvious in the case of deception (where the withholding of relevant information or the conveyance of a falsehood is implied by the word); but it is also true that many instances of manipulation take this form as well. Thus a person manipulates the price of his stock by spreading false rumours of an advantageous merger. We might also say (to use Benson's example again) that women are manipulated by the fashion and beauty industry. A substantial part of our reason for this is that the industry promotes false beliefs

and false values, both about its services and products (using them will make one healthier, happier, etc.) and about their importance in women's lives. Likewise, an employer manipulates her employee when she demands excessive uncompensated overtime from him by leading him to believe that his continued employment depends upon it. While there may be other objectionable features involved in these cases, the element of inducing false beliefs seems vital. My theory entails that when manipulation or deception is utilized so effectively that one could not reasonably expect those who have been subject to either to have avoided falling into error (either evaluative or nonevaluative), then those desires which they approve of on the basis of that error do not thereby have autonomy confered upon them. It would seem, then, that my theory can meet this concern.

The possibility of heteronomy at the second order also led us to consider what we called the regress objection to internalist theories of autonomy which make the approval of a desire sufficient for it autonomy (IV.2). The problem, briefly, is this: on the internalist bi-level theory of autonomy, a first-order desire is autonomous provided the agent identifies with it at the second order. But identification can have this autonomy-conferring status, surely, only if the act of identification is itself autonomous (or if the second-order volition is autonomous). If autonomy depends just upon the attitudes of approval and other psychological states of the agent, then we must determine whether she approves of her identification with the desire itself (or approves of the second-order volition) in order to ensure that it is autonomous. But then we have to determine that this

second act of identification (or one's approval of one's second-order volition) is autonomous by examining one's attitudes toward it, etc. Hence, the regress.

This is a serious objection to internalist theories of autonomy. Is it also an objection to the theory I have proposed here? Insofar as the account of desire autonomy that I offer does not make the autonomy of one's desires solely a function of one's approval of them, it can meet this challenge. For what determines whether a person's approval of a desire confers autonomy upon it are external conditions whose satisfaction can be determined independently of the agent's attitudes or psychological states. Whether one's beliefs and values are false, whether one could reasonably have been expected to avoid falling into error or ignorance, whether one was coerced, and whether one's options were restricted or expanded over a particular period of time can all be determined by independent tests. This is not to say that the subjective attitudes of the agent are irrelevant, of course. An obtuse person could find himself in a situation that would be coercive but for the fact that he did not recognize the coercive threat as a threat, for example, and so the external circumstances involved in coercion may be operative yet fail to actually be coercive because the agent lacks the psychological states involved in coercion. The point is, rather, that there are conditions imposed on autonomous desires that are independent of the psychological states of the agent. If they are satisfied, and the person decisively and unambivalently approves of the desire, then that desire is autonomous. We do not need to consult any higher-order attitudes to determine whether she

approves of her approval of her desire to determine that it is autonomous.

The second worry that Dworkin expressed about what I call internalist theories of autonomy is that they make possible autonomous slaves. Provided that he approves of and identifies with just those desires that his master wants him to have, the happy slave could be autonomous (II.2). I argued that this possibility gives one good reason to reject internalist theories. My theory, though, does not allow this possibility, if (as is surely the standard case) the slave's coming to have these preferences is induced solely by the restricted options he faces or in order to avoid coercive threats. If it is just the external limitations on his freedom of action that induces the slave to revise his preferences so that they conform to those of his master, then this is a case of adaptive preference-revision and it violates the downward monotonicity condition that I argued autonomous desires must satisfy. These same conditions rule out the possibility that the slave's desires are autonomous if they are revised in response to coercive threats issued by his master.

Finally, the case of a person having autonomous yet subservient desires was raised as an objection to the internalist theory of autonomy. Because the subservient or deferential wife could not only have, but decisively and unambivalently identify with, her servile desire to be subservient to her husband, the internalist theory must conclude that her servile desire is autonomous. Insofar as this judgment is implausible, the case succeeds as a counter-example to internalist theories of autonomy. Does this case succeed as a counter-example

to my theory as well? I argued that it does not, because subservience considered in itself limits only the servile person's substantive independence (which is not necessary for the autonomy of one's desires). We think of subservience as problematic because we presume that persons whose moral education was minimally decent and who live in an environment that is not systemically oppressive will not adopt such a posture. Thus, our condemnation of subservience typically depends upon the operation of other objectionable factors, which pervert her beliefs and values (particularly those that pertain to her own worth) in such a way as to undermine the autonomy of the subservient desires she develops as a result of them. Hence, this is not a counter-example to my theory.

Thus my account is able to meet the most important objections which have been raised against bi-level theories of autonomy when they are taken in an internalist sense. It does so, moreover, without embracing the more worrisome features of other externalist theories. It does not make autonomy depend upon respect for one's objective self-interests. Not does it require that one have only true beliefs and values to be self-directed. Insofar as I am able to avoid imposing these as necessary conditions of autonomy, I argued, the theory that I have developed is significantly better than other externalist accounts.

BIBLIOGRAPHY

Arneson, Richard J. (1985). Freedom and Desire. Canadian Journal of Philosophy Vol. 15, No. 3: 425-448.

Babbitt, Susan (1991). Feminism and Rational Interests: Toward a Reconception of Objectivity. Unpublished manuscript; presented at the annual meeting of the Canadian Philosophical Association, May 1991.

Benn, S.I. (1967). Freedom and Persuasion. The Australasian Journal of Philosophy Vol. XLII, No. 161: 259-275.

---- (1976). Freedom, Autonomy and the Concept of a Person. Proceedings of the Aristotelian Society Vol. 66: 109-130.

Benson, John (1983). Who is the Autonomous Man?". Philosophy Vol. 58: 5-17.

Benson, Paul (1987). Freedom and Value. The Journal of Philosophy Vol. LXXXIV, No. 9: 465-487.

---- (1991). Autonomy and Oppressive Socialization. Journal of Social Theory and Practice Vol. 17, No. 3.

Berlin, Isaiah (1969). Two Concepts of Liberty, in Four Essays on Liberty. Oxford: Oxford University Press.

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

Braybrooke, David (1974). From Economics to Aesthetics: The Rectification of Preferences. Nous Vol. 8: 13-24.

---- (1978). Variety among Hierarchies of Preference, in Hooker, Leach, and McClennen, eds., Foundations and Applications of Decision Theory, Vol. 1. Dordrecht, Holland: D. Reidel Publishing Company: 55-65.

Calhoun, Cheshire (1989). Responsibility and Reproach. Ethics Vol. 99: 389-406.

Campbell, Richmond (1979). Self-Love and Self-Respect: A Philosophical Study of Egoism. Ottawa: Canadian Library of Philosophy.

---- (1981). Can Inconsistency Be Reasonable?. Canadian Journal of Philosophy Vol. XI, No. 2: 245-270.

---- (1991). Comments on "Feminism and Rational Interests". Unpublished reply to Babbitt, presented at the annual meeting of the Canadian Philosophical Association, May 1991.

Christman, John (1987). Autonomy: A Defense of the Split-Level Self. Southern Journal of Philosophy Vol. XXV, No. 3: 281-293.

---- (1988). Constructing the Inner Citadel: Recent Work on Autonomy. Ethics. Vol. 99, No. 1: 109-24.

---- ed., (1989). The Inner Citadel: Essays on Individual Autonomy. New York: Oxford University Press.

---- (1991). Liberalism and Individual Positive Freedom. Ethics. Vol. 101: 343-359.

---- (1991). Autonomy and Personal History. Canadian Journal of Philosophy. Vol. 21, No. 1: 1-24.

Darwall, Stephen (1983). Impartial Reason. Ithaca, N.Y.: Cornell University Press.

Dennett, Daniel (1985). Elbow Room: The Varieties of Free Will Worth Wanting. Oxford: Clarendon Press.

Dunn, Robert (1987). The Possibility of Weakness of Will. Indianapolis; Hackett Publishing Co.

Dworkin, Gerald (1970). Acting Freely. Nous Vol. 4, No. 4: 367-383.

---- (1976). Autonomy and Behavior Control. Hastings Center Report 6: 23-28.

---- (1981). The Concept of Autonomy, in R. Haller, ed., Science and Ethics. USA: Rodopi Press.

---- (1988). The Theory and Practice of Autonomy. Cambridge: Cambridge University Press.

Elster, Jon (1979). Ulysses and the Sirens. Cambridge: Cambridge University Press.

---- (1982). Sour Grapes -- Utilitarianism and the Genesis of Wants, in A. Sen and B. Williams, eds., Utilitarianism and Beyond. Cambridge: Cambridge University Press.

---- (1983). Sour Grapes. Cambridge. Cambridge University Press.

Engstrom, Stephen (1988). Conditioned Autonomy. Philosophy and Phenomenological Research Vol. XLVIII, No. 3: 435-453.

Feinberg, Joel (1986). Autonomy, in Harm to Self. New York: Oxford University Press: Chapter 18.

Fisher, John Martin (1987). Responsiveness and Responsibility, in F. Schoeman, ed., Responsibility, Character, and the Emotions. Cambridge: Cambridge University Press: 81-106.

Flanagan, Owen and Amelie Oksenberg Rorty, eds., (1990). Identity, Character, and Morality. Cambridge, Mass; MIT Press.

Frankfurt, Harry (1969). Alternate Possibilities and Moral Responsibility. Journal of Philosophy Vol. LXVI, No. 23: 829-839.

---- (1971). Freedom of the Will and the Concept of a Person. Journal of Philosophy Vol. LXVIII, No. 1: 5-20.

---- (1973). Coercion and Moral Responsibility, in T. Honderich, ed., Essays on Freedom of Action. London: Routledge and Kegan Paul: 65-86.

---- (1975). Three Concepts of Free Action. Proceedings of the Aristotelian Society Vol. 49: 113-125.

---- (1977). Identification and Externality, in A. Rorty, ed., The Identities of Persons. California: University of California Press.

---- (1987). Identification and Wholeheartedness, in F. Schoeman, ed., Responsibility, Character, and the Emotions. Cambridge: Cambridge University Press: 27-45.

---- (1988). The Importance of What We Care About. Cambridge: Cambridge University Press.

Friedman, Marilyn (1986). Autonomy and the Split-Level Self. Southern Journal of Philosophy Vol. XXIV, No. 1: 19-35.

Gauthier, David (1986). Morals by Agreement. New York: Oxford University Press.

Geach, P.T. (1982). Moral Autonomy Still Refuted. Philosophy Vol. 57: 127-129.

Grunbaum, A. (1971). Free Will and the Laws of Human Behavior. American Philosophical Quarterly Vol. VIII: 299-312.

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

Hill, Christopher S. (1984). Watsonian Freedom and Freedom of the Will. Australasian Journal of Philosophy Vol. 62, No. 3: 294-298.

Hill, Thomas Jr. (1973). Servility and Self-Respect. The Monist Vol. 57: 87-104.

---- (1984). Autonomy and Benevolent Lies. Journal of Value Inquiry Vol. 18, No. 4: 251-267.

---- (1986). Weakness of Will and Character. Philosophical Topics Vol. XIV, No. 2: 93-115.

---- (1987). The Importance of Autonomy, in E. Feder Kittay and D. Meyers, eds, Women and Moral Theory. Totowa, N.J.: Rowman and Littlefield: 129-138.

---- (1989). The Kantian Conception of Autonomy, in J. Christman, ed., The Inner Citadel: Essays on Individual Autonomy. New York: Oxford University Press: 91-105.

---- (1991). Autonomy and Self-Respect. Cambridge: Cambridge University Press.

Hoagland, Sarah (1988). Lesbian Ethics: Toward New Value. Palo Alta, Calif.: Institute of Lesbian Studies.

Hurka, Thomas (1987). Why Value Autonomy?. Journal of Social Theory and Practice Vol. 13, No. 3: 361-382.

Jeffrey, Richard (1974). Preference among Preferences. Journal of Philosophy Vol. LXXI, No. 13: 377-391.

Lindley, Richard (1986). Autonomy. London: Macmillan.

Meyers, Diana T. (1986). The Politics of Self-Respect: A Feminist Perspective. Hypatia Vol. 1, No. 1: 83-99.

---- (1987). Personal Autonomy and the Paradox of Feminine Socialization. Journal of Philosophy Vol. LXXXIV, No. 11: 619-628.

---- (1987). The Socialized Individual and Individual Autonomy: An Intersection Between Philosophy and Psychology, in E. Feder Kittay and D. Meyers, eds., Women and Moral Theory. Totowa, N.J.: Rowman and Littlefield: 139-153.

---- (1989). Self, Society and Personal Choice. New York: Columbia University Press.

Neely, Wright (1974). Freedom and Desire. Philosophical Review Vol. LXXXIII, No. 1: 32-54.

Nozick, Robert (1969). Coercion, in S. Morgenbesser et als, eds., Philosophy, Science, and Method: Essays in Honor of Ernest Nagel. New York: St. Martin's Press.

Penelhum, Terence (1979). Human Nature and External Desires. The Monist Vol. 62, No. 3: 304-318.

Rawls, John (1971). A Theory of justice. Mass.: Harvard University Press.

---- (1980). Kantian Constructivism in Moral Theory: Rational and Full Autonomy. Journal of Philosophy Vol. LXXVII, No. 9: 515-579.

Rorty, Amelie, ed. (1976). The Identities of Persons. Berkley, Calif.: University of California Press.

Scoccia, Danny (1987). Autonomy, Want Satisfaction, and the Justification of Liberal Freedoms. Canadian Journal of Philosophy Vol 17, No. 3: 583-602.

Strawson, P.F. (1962). Freedom and Resentment. Proceedings of the British Academy Vol. 48: 1-25.

Taylor, Charles (1976). Responsibility for Self, in A. Rorty, ed., Identities of Persons. Calif.: University of California.

Thalberg, Irving (1978). Hierarchical Analyses of Unfree Action. Canadian Journal of Philosophy Vol. VIII, No. 2.

Thomas, Laurence (1989). Living Morally: A Psychology of Moral Character. Philadelphia: Temple University Press.

Watson, Gary (1975). Free Agency. Journal of Philosophy Vol. LXXII, No. 8: 205-220.

---- ed., (1982). Free Will. Oxford: Oxford University Press.

Wolf, Susan (1987). Sanity and the Metaphysics of Responsibility, in F. Schoeman, ed., Responsibility, Character, and the Emotions. Cambridge: Cambridge University Press: 46-62.

---- (1990). Freedom within Reason. New York: Oxford University Press.

Young, Robert (1986). Personal Autonomy: Beyond Negative and Positive Liberty. New York: St. Martin's Press.