BARRIERS TO SELF-RULE: AUTONOMY, OPPRESSION, & SELF-TRUST

BARRIERS TO SELF-RULE:

AUTONOMY, OPPRESSION, & SELF-TRUST

By

BENJAMIN ELLIOTT WALD, B.A.

A Thesis

Submitted to the School of Graduate Studies

in Partial Fulfillment of the Requirements

for the Degree

Master of Arts

McMaster University

MASTER OF ARTS (2009)	McMaster University
(Philosophy)	Hamilton, Ontario
TITLE:	Barriers to Self-Rule: Autonomy, Oppression, and Self-Trust
AUTHOR:	Benjamin Elliott Wald, B.A.
(McMaster University)	
SUPERVISOR:	Professor Elisabeth Gedge
NUMBER OF PAGES:	

Abstract:

Theories of autonomy are often divided into two broad types: procedural theories and substantive theories. Procedural theories are those that make the criteria for the autonomy of a mental state depend only on the formal features of the mental state, such as its relation to other mental states or features of its etiology. Substantive theories, on the other hand, place some restrictions on what content a mental state can have if it is to be autonomous. Procedural theories have been criticized for failing to explain why those who have internalized oppression are heteronomous, and this has been presented as a motivation for accepting a substantive theory.

In this thesis I dispute this conclusion, and present the outline of a procedural account that can answer this challenge. I argue that the competency account of autonomy suggested by Diana Meyers, in which autonomy is the result of an agent applying a suite of coordinated autonomy skills, provides a new way to understand procedural barriers to autonomy. Using this competency account of autonomy, I show how a lack of self-trust can undermine the successful use of autonomy skills, and thus impede autonomy. I then show how internalizing oppression can undermine self-trust, building on Miranda Fricker's work on epistemic injustice. This shows how procedural accounts of autonomy can account for the heteronomy of the oppressed, undermining one of the criticisms against procedural accounts, and also providing support for a competency account of autonomy.

ACKNOWLEDGEMENTS:

In writing this thesis I have profited greatly from the help of many people. Elisabeth Gedge has been an enormously supportive supervisor. Throughout the entire writing process she has provided encouragement, challenged me to develop my ideas, and suggested literature and lines of inquiry to pursue. My second reader Mark Vorobej has also been invaluable, a meticulous reader who was always ready to point out when my argumentative reach exceeded my grasp. As well as their help in the process of preparing this thesis, I would also like to thank Elisabeth and Mark for their contribution to my development as a philosopher. They have pushed me to develop my skills as a philosopher, introduced me to new ways of thinking and arguing, and greatly enriched my time at McMaster. I would also like to thank the Social Sciences and Humanities Research Council for the CGS master's scholarship that helped support my research.

I also owe thanks to all of the other graduate students at McMaster, who have been immensely supportive and intellectually invigorating. I have learned a great deal from my fellow students through reading groups, seminar discussions, and arguments at the bar. Their company has also kept me sane through the sometimes harrowing process of writing this thesis. I would like to particularly thank Patrick Bondy and Talene Thomasian, whose frequent presence in Wilson's library allowed me to use them as guinea pigs to test out my arguments and intuitions, and who helped me decide on several occasions that I was barking up the wrong tree. I would also like to thank my brother, Joshua Wald, whose complaints about how much less work graduate school was than medical school helped motivate me to put more time into my studies. My parents, Lorne Wald and Deborah Elliott, have provided encouragement throughout my time at McMaster, and a helpful sense of perspective when the stresses of academic life got the better of me. I also want to thank them for being so supportive of my decision to study philosophy. I would also like to thank Jamie Crooks, my undergraduate thesis supervisor, for nurturing my passion for philosophy. Last but certainly not least, I want to thank Ellen Russell for her encouragement and support, and for putting up with all the geographic vicissitudes of my chosen pursuit.

Table of Contents:

INTRODUCTION... 1

CHAPTER ONE: Substantive or Procedural Autonomy...................................... 7

1.1 The Feminist Intuition and the Problem of Oppression.............................. 7

1.2 Substantive Independence and Autonomy.. 12

1.3 Autonomy and Objectivity.. 15

1.4 Autonomy and Knowledge.. 25

1.5 Substantive Accounts and the Desiderata of a Theory of Autonomy.......... 30

CHAPTER TWO: Autonomy and the *ab initio* Requirement.............................. 35

2.1 Introducing the *ab initio* Dilemma.. 35

2.2 The Paradox of Self-Creation.. 38

2.3 The Emergence of Autonomy.. 48

CHAPTER THREE: A Competence Account of Autonomy................................... 58

3.1: Away from Metaphysics, Towards Competencies.. 58

3.2 Barrier to Autonomy from the Deliberative Stance 66

CHAPTER FOUR: Autonomy and Self Trust 74

4.1 Autonomy and "Gas-lighting" 74

4.2 Lack of Self-Trust as a Barrier to Autonomy 87

4.3 Heteronomy, Oppression, and Epistemic Injustice 95

4.4 Conclusion 108

BIBLIOGRAPHY 111

Introduction

When we consider the plight of the oppressed, it is common to judge that one of the harms such individuals suffer is a reduction in autonomy. Sometimes this can be explained through an impairment of the ability of such people to realize their goals. For example, Blacks in South Africa under apartheid were unable to hold political office, and so any Black individual who aspired to politics would have this goal frustrated, and so these people's autonomy was limited in by this external impediment (and many others, of course). However, what about cases in which the oppressed seem not to have goals that are frustrated in this way, and where it seems likely that they lack such goals precisely because they are oppressed?

The kind of case I have in mind is where a pattern of socialization works to foster goals and desires in the oppressed that will not conflict with their continued oppression. For example, women who are raised in highly conservative communities often claim to want nothing more than to be housewives and mothers. Now, such people might be lying or deluded, but it seems plausible that in at least some cases the individuals in question are sincere in claiming not to possess desires that conflict with their assigned social role. In these cases, we still have the feeling that the autonomy of these women is being constrained, despite the absence of any goals that their situation would frustrate. This sort of case is what is sometimes referred to as "false consciousness", but I will not use this terminology in order to avoid begging the question, for the problem is precisely whether and why we should judge this consciousness to be false.

One possible answer would be to say that such people suffer a reduction in their autonomy because they do not have the kind of goals they should have. People ought to want a wider range of options than such people's lives offer them. There are several problems with this type of response, which I will expand on later, but a first stab at the difficulty is that this solution doesn't seem to capture the particularities of the problem it addresses. Lots of people have goals that they shouldn't have, or lack goals they would be better off possessing, and we do not seem to have the intuition that this seriously impedes their autonomy. So what is it about this case in particular that makes us feel that those who are the victims of oppressive socialization are heteronomous?

This is the question I will be attempting to answer in this thesis. One of the first things that becomes clear from looking at the philosophical literature on autonomy is that the term is very vague, and almost certainly used with several distinct meanings. I am focusing on the sense of autonomy that is sometimes referred to as "authenticity." Autonomy in this sense is concerned with the relation between an agent and his or her actions, desires or goals, and denotes a particular kind of "authorship" of the specified action, desire, or goal[1]. This sense of authorship is stronger than the claim that the action was caused by some element of the agent's psychology, or that the desire or goal is a part of the agent's psychology. This extra sense of authorship is often marked by claiming that the actions, desire, or goal is "really" or "deeply" the agent's own.

An example of an impairment of this sense of autonomy is the strong, perhaps even irresistible, desire of a kleptomaniac to steal, despite in many cases firm and longstanding beliefs and values that imply that stealing is wrong and the absence of any reason to desire to steal in the kleptomaniac's current circumstances. In such cases, it seems natural to say that this aberrant desire is less attributable to the individual. One way this is often expressed is by evoking the notion of a "deep self", where those elements that belong to the deep self are what truly constitute the individual, while those elements that conflict with the deep self are external encumbrances on the agent. For example, the desires of the deep self belong to the agent, while those desires that conflict with the deep self just befall the agent in some sense. In a general sense, then, I am concerned with what makes certain actions, desires, or goals truly one's own.

[1] I am using this disjunctive account in order to avoid prejudging any of the substantive philosophical issues in how authenticity should be characterized. It may turn out that authenticity can be characterized in terms of only one of these states, or some combination of them.

The authenticity sense of autonomy is the most suited to addressing our intuitions about the victims of oppression. What we are concerned with in these cases is not whether an individual can do what he or she wants to do, but if what he or she wants to do is what he or she *really* wants to do. As stated, this concern is mysterious, for what can the "really" in the above sentence possibly mean? So the project of this thesis will be to investigate how we can account for our intuition that the victims of oppressive socialization lack the form of autonomy as authenticity that is gestured at above.

I will begin in chapter one by considering Natalie Stoljar's claim that our intuitions about the autonomy of the oppressed cannot be accounted for by any procedural account of autonomy, and that we should therefore look to substantive theories of autonomy. A procedural account of autonomy is one that considers only the formal characteristics and interrelations of the agent's mental system, but places no restrictions on the content of the beliefs, desires, and so forth that make up that mental system. Substantive accounts of autonomy, on the other hand, include some restrictions on the content of an autonomous agent's mental states. If Stoljar is correct that procedural accounts of autonomy cannot explain our intuitions in these cases, then this would be a significant point in favor of accepting a substantive account of autonomy. However, there are significant costs to such a move. I devote the remainder of the first chapter to demonstrating these costs by considering a number of prominent substantive theories of autonomy and raising objections to these ways of understanding autonomy. This critique of substantive theories of autonomy provides motivation for seeking a procedural account of autonomy that can account for our intuitions about the heteronomy of the oppressed.

In the second chapter I investigate a major difficulty for procedural accounts of autonomy in trying to account for our intuition about those with oppressive socialization. Procedural accounts of autonomy claim that the elements of an agent's psychology are rendered autonomous through having the correct relation to other elements of the agent's psychology, for instance second order desires or evaluative judgments. However, it seems that these further psychological elements can only do the job if they are themselves autonomous. This seemingly innocuous requirement, which I call the *ab initio* requirement, leads to a dilemma for procedural accounts. If it is accepted, then it seems that only those who display an implausible degree of independence of their socialization, or even self-creation, will be autonomous. If it is rejected, however, then a procedural account will have to accept that non-autonomous processes or psychological elements can produce autonomy. If this is the case, then how can a procedural account differentiate oppressive socialization from non-oppressive?

This dilemma suggests that a different approach to autonomy is needed. In chapter three, I consider the competency-based theory of autonomy developed by Diana Meyers. This approach makes autonomy an active process of an agent, involving the application of skills of self-discovery, self-definition, and self-direction, rather than a passive property of mental states. This focus on autonomy as an activity opens up new ways of understanding the barriers to autonomy that socialization can impose, and points a way to overcome the dilemma posed by the *ab initio* requirement.

Finally, in chapter four, I investigate the resources this more active conception of autonomy provides for explaining our intuitions about oppressive socialization. In particular, I explain how on a competency-based approach to autonomy we can see that self-trust will play a procedural role in autonomy. By linking this self-trust requirement of autonomy with Miranda Fricker's account of epistemic injustice, we can provide an explanation of how those with oppressive socialization end up lacking self-trust, and thus lacking autonomy. This account allows us to retain both our intuitions about the autonomy of the oppressed and a procedural account of autonomy.

Chapter One: Substantive or Procedural Autonomy?

1.1: The Feminist Intuition and the Problem of Oppression

There is a puzzle in our everyday judgments of who is or fails to be autonomous. In general, we allow that someone's upbringing can have a significant influence on his or her later goals and desires without endangering that person's autonomy. For instance, if we learn that the fact that Bill values helping the poor is the result of his parents' strong emphasis on compassion and empathy for those less fortunate, we are not likely to judge his value heteronomous. However, in some instances learning that a value originated in some sorts of upbringings can cause us to revise our judgment on the autonomy of the value. For instance, if we learn that Judy accepts values that hold that women should be subservient and obey their husbands, and that this is due to an upbringing that presupposed and reinforced these values, we are likely to be much more suspicious of her autonomy.

There are two main strategies we might adopt to explain the difference in our intuitions between these two cases. Firstly, we might appeal to differences in the way that Bill and Judy relate to the values, or differences in the way these values were influenced by upbringing. For example, we might wonder whether Judy's values would survive reflective scrutiny, or whether the values were accepted under conditions of duress. On the other hand, we might appeal directly to the differences in the content of the values endorsed.

These two approaches correspond to the two main varieties of autonomy accounts: procedural accounts of autonomy and substantive accounts. As a first approximation, procedural accounts of autonomy claim that the content of a preference is not relevant to whether or not it is autonomous. Instead, the preference must possess the proper formal features. Examples of the kinds of formal features often picked out include the requirement that the preference be related to one's other preferences in an appropriate way[2], that the preference be responsive to reflective scrutiny[3], or that the individual not experience a feeling of alienation towards the preference[4]. Procedural accounts can be either time-slice theories, in which case the relevant formal properties can be evaluated based on the

[2] For example, Laura Waddell Ekstrom, "A Coherence Theory of Autonomy," *Philosophy and Phenomenological Research* 53, no. 3 (1993)

[3] For instance, Alfred Mele, *Autonomous Agents* (Oxford: Oxford University, 1995)

[4] For instance, Stefaan Cuypers, *Self-Identity and Personal Autonomy*, (Aldershot: Ashgate Publishing limited, 2001) and Diana Meyers, *Self, Society, and Personal Choice* (New York: Columbia University Press, 1989)

agent's psyche considered at a particular moment in time, or historical, in which case the history of the preference or the agent is also a relevant formal feature[5]. One of the upshots of such theories is that, with the right procedural components or pedigree, any preference could potentially be autonomous, and therefore no way of life is ruled out as the possible object of autonomous choice.

Substantive theories of autonomy, on the other hand, include a reference to the content of the preference in determining if the preference is autonomous or not. For example, some substantive theories of autonomy claim that autonomous preferences must be consistent with the agent's future possession of autonomy, or else must match up with the world in some appropriate way. All substantive accounts of autonomy hold that certain ways of life cannot, under any circumstances, be lived autonomously.

Natalie Stoljar claims in her essay "Autonomy and the Feminist Intuition[6]" that no purely procedural account of autonomy will be adequate to account for the ways that oppression impedes autonomy. She explains the

[5] Examples of time-slice procedural accounts of autonomy include Harry Frankfurt, "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68, no.1 (1971), Cuypers, *Self-Identity and Personal Autonomy*, and Keith Lehrer, *Self-Trust: A Study of Reason, Knowledge, and Autonomy*, (Oxford: Oxford University Press, 1997). The most prominent historical account of autonomy is provided by John Christman, "Autonomy and Personal History," *Canadian Journal of Philosophy* 21, no. 1 (1991).
[6] Natalie Stoljar, "Autonomy and the Feminist Intuition," in *Relational Autonomy* (Oxford: Oxford University Press, 2000)

relation between oppression and autonomy in terms of what she refers to as "the feminist intuition", which says that those who internalize oppressive norms are heteronomous. Stoljar doesn't argue for this intuition, but instead takes it as a starting point for her argument. However, I share Stoljar's impression that this intuition has considerable appeal, and thus investigating what kind of theory of autonomy vindicates it is a worthwhile endeavor. Stoljar then proceeds to argue, by reference both to conceptual possibilities and real world examples, that no procedural account of autonomy can vindicate the feminist intuition. This is because oppressive norms cannot be differentiated from non-oppressive norms on the basis of procedural features. Stoljar holds that an oppressive norm can be related to one's other preferences in all the appropriate ways, survive critical scrutiny, and so on, leaving only the fact that the content of the norm is oppressive as grounds for judging the preference heteronomous.

Stoljar describes the feminist intuition as the claim that "preferences influenced by oppressive norms of femininity cannot be autonomous."[7] This formulation seems to be on to something important, but as it stands it is imprecise. In this form the intuition seems to imply that a man's oppression of women would be heteronomous if his actions were to be caused, or even substantially influenced, by his own acceptance of oppressive norms of

[7] Ibid, 95

femininity. This seems like an unintended consequence of the formulation of the intuition; nothing Stoljar says in the article implies that she sees men's oppression of women as being excused in this way. Likewise, although the intuition in this form is restricted to norms of femininity, it seems natural to generalize it to cover all oppressive norms. Therefore, it seems fair to restate the intuition as "agents whose preferences are influenced by norms which are oppressive to that agent are not autonomous."

Neither form of the intuition directly addresses the question of socialization, instead dealing with preference formation. However, it seems clear that the primary source of the internalized oppressive norms, which Stoljar assumes to undermine autonomy, is in early childhood socialization. This is supported by Stoljar's recognition that "the question for all theories of autonomy is what kinds of socialization are incompatible with autonomy."[8] Therefore, the feminist intuition seems to commit us to the claim that children who are inculcated with norms that oppress them are not autonomous, at least in those areas of their life affected by these oppressive norms. Stoljar further claims that no procedural account of autonomy can justify this intuition, and thus we must choose between the appeal of procedural accounts of autonomy and our commitment to the feminist

[8] Ibid, 97

intuition. For Stoljar it is procedural autonomy that comes up short in this contest.

Stoljar uses the example of the women who take "contraceptive risks" as demonstrating the autonomy impeding effects of oppressive norms of femininity. "Contraceptive risks" is the term Luker uses in her Book *Taking Chances*[9] to describe the behavior of women who choose not to use birth control while sexually active, despite not wishing to become pregnant. The explanations offered by these women in interviews are varied, but many seem to be influenced in their deliberation by norms of female sexuality and female assertiveness in relationships that are oppressive to women. Thus, they trigger the feminist intuition that they are heteronomous in these decisions. Furthermore, for Stoljar, we cannot account for this under any procedural theory of autonomy.

Stoljar's claim is that the autonomy impairing effect of the internalization of oppressive norms cannot be accounted for by any formal feature of the preferences so influenced. She provides various arguments for this thesis, tailored to respond to particular accounts of procedural autonomy, but the basic claim can be summed up quite simply. Oppressive norms can be just as consistent, internalized, and identified-with as any other norm. It is only in the specific content of the norm, the fact that it is false and

[9] Kristin Luker, *Taking Chances: Abortion and the Decision not to Contracept,* (Berkley: University of California Press, 1975)

oppressive, that we find a difference between these norms and other more satisfactory norms of preference formation. Thus, a satisfactory theory of autonomy must make reference to the content of agents' preferences.

This is a claim that I wish to refute. However, doing so will take some time. As a first step, I want to establish the problems, both theoretical and practical, with accepting Stoljar's call for substantive theories of autonomy. In order to do so, I will examine several prominent substantive accounts of autonomy, and identify where I take them to go wrong. Doing so will allow me to bring out the theoretical advantages of procedural accounts of autonomy.

1.2: Substantive Independence and Autonomy

The first type of substantive condition on autonomy I will consider is that substantive independence is necessary for autonomy. This is the idea that one's preferences cannot be autonomous if these preferences would restrict one's future autonomy. This condition presupposes a prior specification of what makes a preference autonomous, which could be procedural or could itself contain other substantive elements, and then adds the necessity of preserving one's future autonomy as an extra restriction on what can count as autonomous. For example, it might be claimed that the decision to be a "willing slave" by, for example, joining a religious cult in which one is supposed to defer to the leader's judgment on all matters, or being an obedient housewife who accepts her husband's authority on everything, can never be an autonomous preference. More far-fetched examples might state that one cannot autonomously consent to hypnosis or brainwashing. If there is such a condition, it must be specific to autonomy. It is not generally true that "an action which undermines x is not x". It is perfectly consistent, for instance, to act intentionally in a way that undermines one's own intentionality. An example of this is when someone asks to be put into a medically induced coma. This request is clearly intentional, and the result is the inability to perform intentional actions. The idea that one cannot autonomously choose to be rendered heteronomous would need to be a special case.

When we consider individual examples, it seems implausible that such a condition would apply. While cases where individuals opt for global heteronomy may seem odd, opting for more localized restrictions on autonomy may seem eminently rational, and may in fact be vital to realizing an agent's values and controlling the shape of one's life, and may therefore increase an agent's autonomy rather than threatening it. For example, a smoker may find that her values are frustrated by her inability to quit smoking. She may therefore decide to be hypnotized to remove her desire to smoke. This example may be contested because the agent in question uses heteronomous means to correct an existing heteronomous state. In addition, it may be argued that this substantive condition applies only to cases in which an individual undermines global autonomy, rather than merely undermining local autonomy. In other words, the condition of substantive independence may not apply if an action undermines autonomy in one particular area in order to further one's autonomy overall, but only when one's autonomy as a whole will be substantially impaired in the long-term.

In order to address these worries, let us consider another example. Let us imagine that Bill's values tell him that a life of subservience to God, as a monk, would be the most valuable life he could lead. Of course, on the face of it such a life need not be heteronomous, at least not in terms of values. Bill will not be autonomous in his choice of actions, because these will be dictated by the abbot and by the rules of the order he joins, but his values and preferences may continue to be held autonomously. However, let us further imagine that Bill is drawn to some imaginary order of monks who discourage their members from independent thought, and enforce this through a harsh discipline that over time eliminates the ability to autonomously consider the values currently endorsed. Bill currently possesses the critical faculties required for autonomy, but he can reliably predict that joining this imaginary order of monks will result in him losing these abilities, and becoming heteronomous. This seems like the kind of example that the substantive independence condition would condemn. Bill's decision cannot be autonomous according to this substantive condition.

Now, many of us may wish to say that in forfeiting his autonomy, Bill has made a mistake. If autonomy is something to be valued, he may be choosing to give up a greater value for a lesser one. However, being wrong is no reason to judge someone heteronomous (a claim that will be further defended later in this chapter). The problem with this view is that it seems to require that Bill take autonomy as a value, rather than treating it as a condition of action.

John Christman suggests a similar critique by claiming that the appeal of this sort of substantive condition stems from mistaking two notions of autonomy.[10] On the one hand, there is the notion of autonomy as an ideal we should strive for, and on the other hand there is autonomy as that which denotes a particular kind of authorship of one's actions. As an ideal, autonomy says that all of us should strive to be independent, make up our own minds, live free of coercion, and so on. This ideal may have some value. However, it seems clear that there are other values people may have. It seems illegitimate to rule that a monk who subordinates his judgment to an abbot is somehow not a competent individual, or that such a person's values are not truly their own. Such a life may have its own value, and the monk may

[10] John Christman, "Liberalism, Autonomy, and Self-Transformation," *Social Theory and Practice* 27, no.2 (2001)

honestly judge this kind of life best. To include the ideal of autonomy in the theory of what makes someone's actions their own is to prejudge the issue of what is of value. Even if the ideal of autonomy were the best possible value to hold, this doesn't mean that people couldn't legitimately hold other values, autonomously if not correctly. In developing a theory of autonomy we are looking for what gives certain values or preferences a certain status, the status of being autonomous. While a theory of autonomy should make clear why we tend to think this status has value, it should not presuppose this value within the theory itself. As such, this kind of substantive condition on autonomy is mistaken.

1.3: Autonomy and Objectivity

Another common way to argue for a substantive condition on autonomy is to suggest that we are only autonomous when we are properly guided by the objective facts. Bernard Berofsky presents a defense of this view in *Liberation from Self: A Theory of Personal Autonomy* through the requirement of what he calls "objectivity." Objectivity is a feature of our actions when these actions are guided by the actual features of our situation, rather than being constrained by fixed forms of behavior developed in the past. According to Berofsky, an individual only acts autonomously when his or her action is "open to the world as it is, and is capable of adjusting his [or her] responses appropriately"[11]. This is contrasted to the kind of rigid individual who is stuck in an inflexible way of dealing with the world, unable to adjust to changing situations.

Berofsky illustrates the objectivity condition on autonomy with the example of two brothers, Jean and George, who both become painters due to the influence and encouragement of a third grade teacher, Ms. Webster. Both brothers begin painting out of a desire to please Ms. Webster. However, while the desire to become a painter originates in the same way for both brothers, it is sustained for very different reasons. As Jean develops as a painter, his original motivation falls away and "the character and contours of his adult activity are determined...by contemporary features of this activity."[12] George, on the other hand, continues to paint as a response to his desire to please Ms. Webster. His choice of colors, subjects, and styles are all determined by this original attachment. According to the objectivity requirement, Jean, but not George, is autonomous in his painting.

It is not entirely clear what the source of George's heteronomy is supposed to be. Berofsky suggests that part of the problem with this is a lack of self-knowledge, for "Jean will not change his life if he works through this early relationship [with Ms. Weber] and discards all vestiges of romantic

[11] Bernard Berofsky, *Liberation from Self*, (Cambridge: Cambridge University Press, 1995), 14
[12] Ibid, 201

feeling for the teacher. George, on the other hand, may be profoundly affected by a significant change in the traces left by this relationship...There is a fragility to George's career."[13] This suggests that the real problem with George may not be that his motivation is backward looking, but merely that he would disown this motivation if he were aware of it. If this is in fact the point of the objectivity criterion, then it is actually a straightforwardly procedural requirement, and in fact one that will make up a part of what I take to be the most plausible theory of autonomy, and so presents no problem. However, this doesn't seem to square with the earlier talk of being open to the world and flexible, since someone could conceivably have a rigid view that was closed to the world, but still endorse this view upon reflection, even once aware of the sources of this view.

If the fragility of George's lifestyle is not the central point, then the criticism may be that his desires are caused by a "backward-looking" fixation, rather than responding adequately to the situation he finds himself in. It is not quite clear to me how this backward-looking fixation is supposed to work. After all, the adult George is still in some ways responding to the features of the actual situation he finds himself in. It is a fact that the adult George has unresolved romantic feelings for Ms. Webster, that he feels an affinity for painting with certain colors because Ms. Webster liked them, and

[13] Ibid, 202

so on. We don't necessarily need to include any references to past states of affairs to account for his behavior. Perhaps we could interpret George's situation as one where, as an adult, he is painting in order to please Ms. Webster as she was when he was in the third grade, and thus his current actions are aimed at a goal that has since ceased to exist, since the past version of Ms. Webster that George is painting for no longer exists. But this seems to collapse back into a case of self-deception that violates purely procedural requirements of autonomy. If George were aware of the goal that his painting has aimed at, then he would either give up the goal, thus showing its fragility to reflection, or else keep the goal. Assuming that George has fairly standard beliefs about the (im)possibility of pleasing past versions of people, he would then be accepting incoherence between his beliefs and his goals, and this would violate the minimal rationality requirement that is a component of most procedural accounts of autonomy.

However, even if we could make sense of a backward looking fixation, it is not clear that this would actually be a problem for autonomy. Let us imagine that, instead of being influenced by a third grade teacher, George became a painter because his mother had always wanted an artist in the family, and he had promised her on her deathbed that he would become a painter in order to fulfill this dream of hers. This motivation is every bit as backward looking, rigid, and unresponsive to the world as was the desire to please one's third grade teacher, but it is not at all clear to me that in this case we should judge George heteronomous. We may debate the moral status of deathbed promises, or the rationality of George's actions, but to judge it impossible to take on such a promise autonomously seems implausible. In general, having one's motivations be backward-looking seems to be a condition of many kinds of commitment, and to rule them all heteronomous seems far too counter-intuitive a result to countenance.

If it is not the fact that George's motivations are backward looking *per se* that renders him heteronomous, the only option left is that it is the fact that his motivations do not *in fact* provide good reason to be a painter. Trying to please a former third-grade teacher just does not give one a good reason to choose a career. This will also account for the fragility of this goal under reflection; goals supported by poor reasons, or no reasons at all, are clearly more vulnerable to being revised upon reflection than are those supported by good reasons, other things being equal. If this is the reason for George's heteronomy, then the objectivity criterion boils down to the requirement that we hold the goals and values that our reasons best support, and that we hold them for these reasons. This may also be the kind of condition that Stoljar is aiming at. One of the motivations she offers for the feminist intuition is that "women who accept the norm that pregnancy and motherhood increase their worthiness accept something *false.*"[14]

While objectivity in this sense is clearly a highly desirable feature, I would deny that it is a feature of autonomy. To include the objectivity criterion as a condition of autonomy is to conflate two very different abilities. Autonomy is intended to pick out the power to author one's own actions, to take responsibility for what one does and the values one accepts. The ability to get things right appears to pick out another power, one which Paul Benson refers to as "orthonomy" or right rule, as opposed to self-rule[15]. In general, it seems that these abilities describe separate powers. One could be forced to believe the truth in a way that bypasses autonomy, and it seems strange on the face of it to claim that the opposite is not also possible. The plausibility of the objectivity condition on autonomy may derive its force from the desirability of orthonomy. No matter how desirable orthonomy may be, however, it is still a mistake to include it as a condition of another state altogether, that of being self-ruled.

In support of this contention, consider the situation of agents in a world in which a Cartesian evil demon is operating. Let us imagine that these agents are psychologically identical to us, but that because of the systematic

[14] Stoljar, "Autonomy and the Feminist Intuition," 109

[15] Paul Benson, "Feminist Intuitions and the Normative Substance of Autonomy," in *Personal Autonomy* (Cambridge: Cambridge University Press, 2005), 132

deception of the evil demon none of their beliefs or values correspond to reality. This would clearly undermine the possibility of orthonomy, but it is not clear why this would alter the agent's ability to govern his or her own actions. If true, this implies that orthonomy and autonomy are separate abilities, and the conditions for each should be kept separate.

In "Sanity and the Metaphysics of Responsibility", Susan Wolf presents a view of responsibility which embodies a substantive requirement similar to Berofsky's, which Wolf calls the "sane deep self view"[16]. This view is presented as an account of responsibility, and so it may seem misleading to discuss it as a theory of autonomy. However, Wolf compares her theory to the theories of autonomy espoused by Frankfurt, Watson, and Taylor, so there is clearly some connection between Wolfs view and the theory of autonomy. Even if Wolf herself would not characterize her view as a theory of autonomy, it would be easy enough for a supporter of substantive theories of autonomy to adopt her criterion of responsibility as a criterion of autonomy. Therefore, I shall consider the suitability of Wolfs "sane deep self view" as specifying a substantive condition on autonomy. Like the objectivity requirement, the sane deep self view holds that being in the correct connection to the world is essential for autonomy. However, for the sane deep self view what is necessary is not that we be in fact guided by the

[16] Susan Wolf, "Sanity and the Metaphysics of Responsibility," in *Personal Autonomy* (Cambridge: Cambridge University Press, 2005)

correct reasons that pertain to our decision, but instead that we have the capacity to be so guided, whether or not it is actualized in any given decision. Let us see if this slightly weaker substantive condition is more plausible.

Wolf begins by describing what she calls the "deep self view" of autonomy, a kind of autonomy account that for Wolf includes the views of Frankfurt, Watson, and Taylor. The essence of this view is the idea that to be autonomous one's actions must stem from one's "deep self." The deep self has been variously characterized as one's second order desires, for Frankfurt, one's values, for Watson, or one's reflective and critical faculties, for Taylor. In each case, what is crucial is that autonomy, according to these versions of the deep-self view, requires that some further level of the self be involved in making decisions.

Wolf believes that all of these views fall prey to the same criticism, a criticism that has marked similarities to Stoljar's feminist intuition. She asks us to consider the thought experiment of Jojo, son of an evil dictator. Jojo is raised to follow in his father's footsteps, and because of his twisted upbringing he comes to act in the same way as his father. Furthermore, he wants to want to act in the evil ways that he acts, and he values this kind of life, and he may even have intact reflective and critical faculties that he uses to justify and pursue his evil lifestyle. In other words, Jojo's deep self appears to have been shaped by his upbringing so as to support his evil lifestyle. Wolf claims that there is an intuition that, because of his warped upbringing, Jojo is not responsible for his evil actions. We might plausibly also question whether Jojo is autonomous. Wolf locates the source of this intuition in the fact that, given his upbringing, Jojo could not have turned out differently than he did. He has been rendered unable to respond to the correct moral reasons. His twisted upbringing has resulted in a kind of moral blindness. Wolf therefore holds that autonomy requires the substantive condition that a person retains the ability to recognize and act on the correct moral values.

Wolf refers to this condition as the "sane deep self" requirement. We often judge the insane to be heteronomous, and Wolf claims that this is so because the insane lack the ability to know whether what they are doing is right or wrong, and thus cannot be responsible for acting wrongly. But Jojo also lacks the ability to tell that what he is doing is wrong, and has been rendered unable to do so by his warped upbringing. Thus, Jojo is in a sense insane because of his inability to be responsive to the proper values. Wolf admits that this use of the term insane does not fit common usage, and imports strong normative conditions on sanity. However, she justifies retaining the term sanity on the grounds that we deny the responsibility of the insane for the same reason that we should deny the responsibility of people like Jojo, and thus there is enough continuity between this special sense of sanity and the ordinary usage to justify retaining the term. This substantive condition on autonomy allows us to justify our intuition that people raised with oppressive values are not autonomous without judging ourselves heteronomous at the same time, or requiring some kind of metaphysically dubious self-creation.

However, this solution has several serious faults that militate against our accepting such a strongly normative "sanity" condition on autonomy. If interpreted as a condition on autonomy, Wolf's sane deep self view seems to be vulnerable to the same criticism as the objectivity condition, namely that it confuses autonomy with orthonomy. This argument might seem less forceful in this case since Wolf's theory only requires the possibility of acting according to the correct moral values, not that the agent actually act that way in all cases. However, this still seems too strong. The inhabitants of a Cartesian demon world could still hold their values autonomously, even if the demon had set up the world so as to make it impossible to perceive the correct moral values. Likewise, unless we are very optimistic about people's trans-historical ability to discover the correct moral values, this will mean that most people in ancient times would be heteronomous, since they would be unable to believe that slavery and discrimination were wrong. Likewise, it would seem to render our own possibilities for autonomy rather dubious, unless we think that our time in history is the very first period to be free of deeply entrenched moral mistakes.

One way to avoid these unwelcome consequences would be to restrict what kind of inability to respond to proper values we count as undermining autonomy. We might say that there is a difference between Jojo and our hapless evil demon victim in that the evil demon victim retains the ability to recognize right from wrong, but is prevented from realizing it by his unfortunate circumstance, whereas Jojo is unable to recognize right from wrong due to an internal incapacity, one that would follow him in any close counter-factual situation. I don't think such an approach can work, however. Wolf herself does not seem to accept such a distinction, since she explicitly accepts that people in past generations would turn out not to be responsible for certain kinds of injustices on her criterion (and hence not autonomous for the autonomy version of her criterion), since their situations prevented them from responding to the correct values.

In addition, interpreting the condition in this way seems to transform it into a procedural requirement, instead of a substantive one. After all, it seems that actually bearing the proper relation to the external features of the world, such as correct values, is no longer doing any of the work. Instead, what is important is that the agent is so constituted as to be able to identify correct values if the external environment is suitably cooperative. But being so constituted is just a matter of the arrangement of the individual's psyche, and thus would involve only procedural criteria.

Wolf responds to the criticism that her criterion implausibly judges past generations free of responsibility for certain injustices by saying that we always need to make our judgments of responsibility (and presumably autonomy as well) from where we are now, and thus make use of our best current understanding of what the correct moral values are. However, this response seems to lose some of its plausibility if we challenge what counts as "our" current understanding of moral values. Different cultures within our own society and around the world will often have divergent conceptions of what counts as the correct moral values. Thus, Wolf's advice to make our judgments based on our present understanding of moral values would seem to render people in other cultures heteronomous as a group. I take this to be an extremely disturbing result, especially given the place of autonomy in defending minorities from paternalistic state intervention. This suggests that Wolf's criterion, if interpreted as a theory of autonomy, fares no better than Berofsky's objectivity criterion.[17]

1.4: Autonomy and Knowledge

A slightly weaker version of the objectivity requirement on autonomy is what I shall call the "knowledge requirement." This is the idea that an

[17] One lesson to take from the failure of this Wolf style account of autonomy is that the link between autonomy and responsibility may not be that direct. This would explain why an attempt to convert a theory of responsibility into a theory of autonomy is unsuccessful; a similar issue emerges in the next section.

agent must have access to the relevant information in deliberating on what preferences or values to accept. Alfred Mele presents a version of this requirement through the thought experiment of King George.[18] King George is a most unfortunate individual, who desires to do good for the subjects of his kingdom but is cursed with malicious advisors. These advisors systematically mislead King George as to the conditions of his kingdom and the results of his policies. Since King George gets all of his information from these advisors, he bases all of his policies and decisions on this erroneous data. The result of this is that every policy enacted has precisely the opposite of its intended effect, and the more he tries to do good the more miserable his subjects actually are. It seems unfair to say that king George has autonomously rendered his subjects miserable, and this prompts the intuition that there is a requirement that an agent have some level of knowledge in order to qualify as autonomous.

The thought experiment as presented by Mele concerns the autonomy of actions rather than preferences, but it is easily adapted. Our preferences may not always depend on knowledge to the same extent as our actions do, but there is still a significant influence. For example, my preference to avoid pain may be based on very little knowledge, needing only a very few experiences of pain in order to form it, but my further preference for

[18] Mele, *Autonomous Agents*

avoiding contact with live electrical wires is based on my belief that such contact will cause pain. Likewise, a woman who has been raised under oppressive patriarchal values may come to accept them because she has been told, and believes, that studies have shown women are less intelligent and have less self-control than men. Thus, false beliefs may have a strong influence on our preferences, and this may be considered sufficient to undermine the autonomy of these preferences.

Of course, a lot rests on what information we count as relevant to the formation of the preference. For instance, it seems as if the knowledge as to whether a given value is good or not is clearly relevant to our holding it, but if we allow this to count then the knowledge requirement will collapse into the objectivity requirement, whose deficiencies were explored above. In order to avoid this, it seems sensible to restrict the relevant information to information about facts, rather than about values[19]. Why should we believe that the lack of this kind of information impairs autonomy? One consideration is that autonomy is often considered a condition of responsibility. If someone was acting autonomously, it seems right to hold them accountable for their actions, and thus such an agent may merit praise or blame. However, if someone does wrong due to (non-culpable) ignorance

[19] The fact value distinction is much less clear-cut than I suggest here, and there may well be boundary cases. Nonetheless, I will assume that we can make this distinction, at least in paradigmatic cases, well enough to give this suggestion some plausibility.

we do not normally feel that they deserve blame. Likewise, in the less frequent opposite case, we may feel that the right sort of ignorance can remove praise from an otherwise virtuous agent. For example, imagine that Ted wishes to save Sally from a snakebite by administering the antivenin. He has good intentions, but in his ignorance he fails to realize that the antivenin he administers is the wrong variety for the type of snakebite Sally has received. Luckily, the bottle was mislabeled and Sally's life is saved, but we might still judge that Ted's ignorance reduces the amount of praise he is due.

While I agree with these intuitions about responsibility, it seems a mistake to try to explain these intuitions by claiming that what the individuals in question lack is autonomy. This mistake originates in the temptation to make autonomy the sole condition of responsibility, in which case all failures of responsibility would be failures of autonomy. However, autonomy denotes a specific kind of ownership of one's actions. This is plausibly a necessary condition for responsibility, but I see no reason to assume it is a sufficient condition. Instead, we should agree that ignorance could influence responsibility, but deny that this influence operates via autonomy. Actions done in ignorance are no less one's own actions, and values held due to mistaken beliefs can still be genuinely one's own, even if one has a valid excuse for holding them.

In order to further motivate the view that ignorance should not be held to impede autonomy, I would like to draw attention to some strange consequences this view would imply. The example of king George posits human malevolence as the source of ignorance, but this risks distorting our intuitions. If we take a case of "natural" ignorance, I think our response is likely to be more ambiguous. For instance, given the advances in knowledge through time, this view would seem (again) to render past generations largely, if not entirely, heteronomous. After all, in many areas of life past generations seem to be as systematically ignorant as king George. By extension, it seems likely that we also lack important information we might one day possess. For instance, many supporters of capital punishment do so at least in part due to a belief that this form of punishment creates a deterrence effect, while many opponents of capital punishment reject the idea that executions serve as a deterrent to crime. Psychological evidence for this claim is currently contentious, but presumably it could become decisive, at which point the preferences of all those currently involved in the debate would be revealed to have been heteronomous if we accept a knowledge requirement on autonomy. This would threaten the existence of autonomy. It also lacks the intuitive support that the king George example elicited, at least to my mind.

Of course, if we are systematically deceived, either by nature, a Cartesian demon, or the maliciousness of other humans, then we will often fail to achieve our aims. This is certainly the fate of poor King George, and it is possible that this could be true of our preferences or values as well as our actions. However, in the same way we distinguished between orthonomy and autonomy earlier, we should likewise distinguish between efficacy and autonomy. Without relevant information, we may come to regret the values we currently endorse, but this unfortunate possibility does not make our present endorsement any less valid. In general, if we focus on the self-ownership role of autonomy, it is mysterious why the fact that I lack knowledge would make my preferences or values less my own. Some level of ignorance is an inevitable feature of the human condition, and we must be able to formulate and endorse preferences despite this if autonomy is going to be practically attainable.

1.5: Substantive Accounts and the Desiderata of a Theory of Autonomy

I hope to have cast doubt on what I take to be the most plausible substantive accounts of autonomy on offer. Given the constraints of space, I have not been able to address every substantive theory that has been developed, and even if this were possible it would not address the possibility that some substantive theory yet to be developed would be the best account of autonomy. Therefore, in order to lend further support to my claim that we must turn to a procedural account of autonomy, I wish to consider two more general features of autonomy accounts that I take to suggest the superiority of procedural views of autonomy.

The first consideration for an account of autonomy is that autonomy should be a characteristic of *agents*[20]. Autonomy is a term of evaluation applied to an agent and his or her relation to his or her preferences and values. If we accept this, then it implies that changes in the external environment that the agent is not aware of should not cause us to change our evaluation of the autonomy of an agent. For example, if an autonomous agent is snatched from our world into a world controlled by a Cartesian demon, but the demon ensures that agent's experience is unaltered, this should not affect our evaluations of the agent's autonomy. After all, nothing within the agent has changed, either directly or through contact with the changed external environment, so why should the presence of any features that describe the agent be altered? This is not to say that changes in the external environment cannot over time act to erode or enhance autonomy. However, this is accomplished through a change in the agent, a change that is brought about by the influence of the environment over time. Autonomy tells us something

[20] This restriction applies only to autonomy as authenticity, as I delineated it in my introduction. There are some senses of autonomy that do not focus on agency exclusively, for instance the way in which someone who is denied a job on the basis of prejudice has had their autonomy restricted, but the autonomy that is concerned with the "deep self" is a characteristic of agents.

about the ownership of values and preferences, and I do not believe that a change in external elements of the world, independent of any influence on the agent, can tear away our values.

In support of this proposed constraint on autonomy accounts, we should notice that if we were to accept that the environment needs to be a certain way in order to permit autonomy, we might discover that our own environment fails to meet these criteria. If we were to discover this, it would mean that we had been heteronomous all along. It would also run the risk of tailoring autonomy too closely to the modern, and perhaps western, environment. If other time periods, and other cultures, lack the proper environment then we would judge that, while they may have believed that they autonomously endorsed and held to their values and preferences, in fact these values were not authentically their own, unlike ours. This seems like a patronizing and implausible result. Or, of course, it might turn out that some other culture but not our own possesses the requisite external features, which would be less patronizing but no less implausible. It is possible that our culture, or even all cultures, are characterized by massive and pervasive heteronomy, but such a conclusion would require a great deal of support, placing the burden of proof firmly on the proponent of such a theory.

I wish to note that this condition on autonomy may at first glance seem to rule out not just certain substantive accounts, such as those that endorse an information or objectivity condition on autonomy, but also to rule out any reference to the historical conditions that led to the agent's current psychological state, such as John Christman's account of autonomy. Christman's account makes autonomy a feature of an agent's history. A person's preferences are autonomous for Christman only when the individual would not have resisted the process whereby the desire was formed had they been aware of this process as it was taking place[21]. Indeed, Christman argues that any plausible account of autonomy must include a historical component, for otherwise we will be unable to distinguish between two agents, one of whom has developed normally up to that point, while the other has just moments ago had his or her entire psychology invasively rearranged by a nefarious mind-control ray, or similar external intervention[22]. It might seem that this account locates autonomy somewhere external to the subject, namely in historical events, and thus would violate the methodological constraint we are considering. This would be a problem, since I am concerned to rule out substantive theories at this stage, and Christman's account is considered, by himself and others, to be a procedural account, so if it were caught in this net it might suggest that the criterion being used was too restrictive. I do not, however, believe that Christman's theory, or any other theory that references the agent's history, is actually

[21] Christman, "Autonomy and Personal History"
[22] Ibid

vulnerable to this particular criticism. Instead, historical accounts of autonomy imply that an agent is a temporally extended phenomenon. Earlier time-slices of a person remain a part of the individual as a whole, and thus the historical non-resistance condition on autonomy, and other historical conditions like it, can be seen to refer back to things that are still internal to the individual in question.

A second requirement stems from what I take to be the primary practical value of theories of autonomy, which I take to be a feature that speaks in favor of a "right to do wrong". Autonomy is appealed to as a consideration in favor of allowing individuals to live by their own lights, according to their own values, even when we might disagree with them. Of course, the theory of autonomy in itself cannot dictate how we should weigh the importance of allowing people to live according to their own values, or how we should balance this against duties of beneficence and non-maleficence. However, such a theory should be able to be appealed to in this debate in order to indicate under what conditions the importance of respecting an individual's own choice as to how to live kicks in.

This requirement clearly rules out theories along the lines of the objectivity criterion discussed earlier. While there is no (conceptual) problem with claiming that we should not value autonomy but instead only value getting things right, this should not claim to be a theory of *autonomy*. It is more properly an argument for the unimportance of autonomy, but this claim must itself be made in reference to a theory of autonomy that does take seriously some conditions under which a person can be autonomous and wrong.

If we reject substantive accounts of autonomy, then we are left to try to account for why those raised in oppressive contexts are often thought to be heteronomous. Unfortunately, as we shall see, accounting for the origins of autonomy is a particular problem for procedural accounts of autonomy. We must now consider this problem, and see which procedural theory of autonomy will allow us to explain the feminist intuition.

Chapter Two: Autonomy and the *ab initio* Requirement

2.1: Introducing the *ab initio* Dilemma:

Procedural accounts of autonomy are caught in a particularly difficult double bind when it comes to the question of socialization. The seeds of this difficulty can be identified in John Christman's innocuous-seeming objection to Frankfurt's theory of autonomy, which he calls the *ab initio* problem. Frankfurt's theory states (very roughly) that our first order desires are autonomous if they are matched with our second order desires. In other words, my desire to study philosophy is autonomous if I also desire to desire to study philosophy. Christman objects that nothing in this account certifies the autonomy of the second order desires. Frankfurt's account seems to imply that if an evil scientist uses a mind-control ray to alter both my first order desires and my second order desires, then my autonomy has not been infringed upon. If our second order desires are not autonomous, then it seems mysterious how these second order desires could secure the autonomy of our first order desires. Of course, we could give a Frankfurt style account of the autonomy of second order desires by appealing to third order desires, but this doesn't seem to help our dilemma, since the same problem reoccurs at this level, and appealing to an infinite hierarchy of desires is both empirically and conceptually suspect.

The problem seems to be that if we want to certify that a certain desire is an expression of self-rule, then whatever is doing the certifying had better be an expression of self-rule itself. This suggests a general condition on procedural accounts of autonomy, which I shall follow Robert Noggle in calling the *ab initio* requirement[23]. The *ab initio* requirement holds that if a theory is going to explain autonomy by reference to features of an agent's psychology, as any procedural theory will have to, then these further elements must themselves be autonomous. This requirement has intuitive appeal, but unfortunately it leads us to a dilemma.

If we accept the *ab initio* requirement, then it appears that in attempting to certify that a given preference is autonomous we have only two options; either we must appeal to some other autonomous element of the agent's psyche that renders it autonomous, or else we must claim that its autonomy is self-generated, i.e. the preference itself secures its own autonomy. Of course, if we choose the first strategy we will then be

[23] Robert Noggle, "Autonomy and the Paradox of Self-Creation: Infinite Regress, Finite Selves, and the Limits of Authenticity," in *Personal Autonomy* (Cambridge: Cambridge University Press, 2005)

challenged to account for the autonomy of this further element, and so we will need to hold that some part of the agent's psychology is able to supply its own autonomy, or else hold that autonomy is some kind of holistic property that is created by the mutual support of multiple elements in the agent's psychology. Since autonomy refers to self-rule, the most obvious way to do this would be to claim that the autonomy of preferences is self-created when these preferences are freely chosen, i.e. they are generated by an act of contra-causal free choice, or that autonomous preferences emerge from a psychological system that is founded on elements that have originated in this way.

The first problem with this is that it seems to render autonomous decisions groundless, since if one were to adopt a preference for specific reasons, then these reasons would seem to be determining one's preferences. Since reasons are not the sorts of thing that can be autonomous, the preference in question would fail the *ab initio* requirement. Furthermore, contra-causal freedom has been criticized as metaphysically dubious, if not incoherent, and so making such freedom essential to autonomy seems to endanger the possibility of autonomy. However, even if these problems could be overcome, or accepted, the contra-causal account would undermine our attempt to account for the feminist intuition. If autonomous preferences must be free of all social determination, oppressive or otherwise, then it seems that someone raised under oppression would be no worse off than anyone else when it comes to autonomy.

We could, of course, reject the *ab initio* requirement. We might claim that there is nothing mysterious about autonomy emerging from non-autonomous elements; any more than there is something mysterious about life originating from non-living matter. In this case the fact that things that are not autonomous have influenced our preferences is not taken to be problematic, at least not in and of itself. Thus, we can accept that our preferences will be influenced by our early childhood socialization without fearing that this renders us heteronomous. However, if we take this route we will need to find a way to distinguish oppressive socialization from those kinds of socialization that are taken to be amenable to autonomy. Many of the autonomy theories that reject the *ab initio* requirement do not succeed in doing so, and thus they suggest that the victim of oppression is in fact autonomous in later life, contrary to the feminist intuition. Thus, procedural accounts of autonomy seem to face a problem in accounting for the effects of socialization. I will start by exploring some theories on both sides of the *ab initio* divide in order to clarify the problem we face.

2.2: The Paradox of Self-Creation

One objection to my characterization of the *ab initio* requirement might be that Christman, who introduced the objection, holds a procedural account of autonomy and does not think that his view implies self-creation. Indeed, he explicitly affirms that "no person is self-made in the sense of being a fully formed and intact 'will' blossoming from nowhere."[24] However, I believe that Christman's historical account of autonomy does in fact entail that autonomous individual's must demonstrate an implausible independence of their socialization and environment, an independence that would seem to require that the individual be self-created. To see why, let us examine Christman's theory more closely.

The core of Christman's theory is captured in three conditions,

> "(i) A person P is autonomous relative to some desire D if it is the case that P did not resist the development of D when attending to this process of development, or P *would not have* resisted that development had P attended to the process;
>
> (ii) The lack of resistance to the development of D did not take place (or would not have) under the influence of factors that inhibit self-reflection;
>
> and (iii) The self-reflection involved in condition (i) is (minimally) rational and involves no self-deception."[25]

[24] Christman, "Autonomy and Personal History," 1
[25] Ibid, 11

Central to Christman's account is that the autonomy of a desire rests on the process by which the desire came about, rather than how the desire fits into the current "time-slice" of the agent's psychology. For example, if Raymond is hypnotized into having a desire to assassinate the president, then presumably had he been in a position to reflect on the process by which this desire was formed he would have resisted it. Thus, it is not any feature of the desire itself that renders it heteronomous, but the fact that it was imposed on Raymond through a mechanism that he would resist, were he to have attended to it. We need the counter-factual here because part of the process of hypnosis is presumably to render the subject unaware of the process that is taking place, and in less extreme cases we may simply fail to notice the processes that are forming desires in us, because they are subtle and gradual, because of inattention, due to ignorance as to the effects of the relevant processes on one's desires, and so on.

Condition (ii) is necessary in order to prevent a situation in which someone does attend to the process by which a desire is formed, and doesn't resist it, but this lack of resistance is due to external pressures that serve to inhibit the ability to decide whether or not to resist the process in question. For instance, we can imagine the example of someone who joins a cult and forms the desire to obey the dictates of the cult through the cult leader's speeches. The individual in this case may well have been aware of the source of this new desire, and not have resisted it. However, in this case the lack of resistance is due to the fact that part of the indoctrination procedure involves sleep deprivation and fasting, which has clouded the individual's reasoning ability. Thus, we need to say that the reflection occurs free of 'reflection-constraining factors.'[26]

Particularly relevant to our present investigation is the fact that Christman identifies certain education techniques as including reflection-constraining factors when these education techniques undermine, rather than enhance, the student's ability to critically reflect on what they have learned. Presumably, the same would apply to various types of socialization, and this seems to suggest a way to substantiate our judgment that oppressive socialization can produce heteronomy. This seems promising for our purposes; however, when we press on his account it soon runs into trouble.

Christman's account of autonomy rests on the agent's own judgment as to the acceptability of the type of desire forming process in question. It is not any intrinsic feature of certain types of desire forming processes that makes the resulting desire heteronomous, but just the fact that the agent in question would resist that process in the circumstances in question. This is shown by the fact that one and the same desire forming process, say hypnosis, could produce autonomous desires in one case, where the agent

[26] Ibid

chooses to be hypnotized to stop smoking, and heteronomous in a another, when the agent is hypnotized to kill the president. Thus, it is the agent's judgment that is the key factor. But on what grounds does the agent make this judgment?

It seems clear that the agent must make the judgment as to whether to resist the desire forming process in question based on his or her existing stock of beliefs and desires. However, according to the terms of the *ab initio* requirement, in order for these desires and beliefs to issue in a judgment that secures the autonomy of a further desire, they themselves must be autonomous. In order to assess the autonomy of these desires we will need to look at the process by which they were formed, and whether the agent did or would have resisted this process. This will inevitably lead us further and further back into the subject's history, until we arrive at a point where it makes no sense to say that the individual could resist or fail to resist the formation of desires, because the individual in question will be too young to make evaluative judgments of desire forming mechanisms. At this point it becomes unclear what it means to ask if the individual would have resisted the process of belief formation if he or she had attended to it. What grounds for decision are we to imagine the child appealing to in deciding whether or not to resist this process?

Clearly we cannot use the standards the child has at the time the decision must be made, for there are no such standards. Likewise, there is no fact of the matter about what the individual's standards would have been without any process of socialization, for human beings just do not develop without forms of external input. If we appeal to the standards the child will have after the process of socialization is complete then the process of socialization will itself influence the judgment of its own acceptability.

It is perfectly possible that a process of socialization will result in an agent who judges that very process illegitimate. We could conceivably use indoctrination and brain-washing to instill the value of independent thinking, on the basis of which our own process of socialization could be criticized. However, it is also possible to use socialization to instill a set of standards that cannot be turned against the process whereby these standards were introduced. For instance, it is conceivable that a socialization process that involved brainwashing could be used to inculcate a set of values that revolve entirely around obedience to God, and the view that brainwashing is a legitimate process of desire formation as long as it led to proper obedience to God. Thus, it would seem that a person raised this way would judge his or her own upbringing as something that should not be resisted, and that a large part of the explanation of this judgment would be the nature of the individual's upbringing. This seems contrary to the *ab initio* requirement, and also appears to allow for oppressed individuals to autonomously endorse the values and desires that oppression have inculcated, contrary to the feminist intuition.

The only other alternative, and the one Christman himself endorses, is that we can make use of the individual's own later acceptance of his or her own process of socialization, provided that this "acceptance is not simply the result of those processes [of socialization] themselves, especially if they are such that reflective, retrospective, evaluation is distorted by them or made impossible."[27] So we can rely on an individual's own later evaluation of his or her own socialization process, as long as this evaluation is itself independent of the socialization process. But how is such independence possible?

Here's one possibility. Perhaps one can use one's beliefs to evaluate and criticize one's process of desire formation, and thus judge some desires heteronomous. Many of our beliefs are clearly formed in a way that is largely independent of our socialization. After all, the pressures of the outside world will tend to influence beliefs in ways independent of socialization. Thus, if beliefs on their own are capable of providing a basis for an individual to satisfy Christman's criteria for autonomy, he might seem to have succeeded

[27] John Christman, "Autonomy, History, and the Subject of Justice," *Social Theory and Practice* 33, no. 1 (2007), 10

in avoiding a regress[28]. However, I do not think that beliefs on their own can do the job.

To see why this is so, consider the kind of belief to which we would need to appeal. It might seem that we can appeal directly to a belief with a content like "socialization process x ought to be resisted". However, what would ground such a belief? It could not be the fact that the process of socialization in question is heteronomous, since this is what the judgment itself establishes, and thus cannot be presupposed. Perhaps it could be grounded by a further clause, such as "socialization process x ought to be resisted because it involves punishing a child for independent thinking." This does not yet get us what we need, however, for we do not know why it is that the individual thinks that punishing a child for independent thinking entails that such socialization ought to be resisted. What sort of reason could the individual have for making this judgment? There are two possibilities here; either the reason is based on the individual's desires, or it is independent of the individual's desires.

If the reason is based on the individual's desires, for example "I ought to resist any process of socialization that depends on punishing people for independent thinking because I do not want my desires formed in such a

[28] Whether or not beliefs formed in this way would themselves satisfy the *ab initio* requirement, influenced as they are by the facts of the external world and the norms of epistemic rationality, is another question, but one I will not attempt to address.

way", then the appeal to beliefs collapses back into an appeal to desires, and we are back where we started. What if the reason is independent of desires? This too runs into problems. The reason for holding a belief about the acceptability of a given kind of socialization cannot be the objective fact that one ought not to accept that kind of socialization. Even if one accepts the existence of such normative facts, a contentious position, if Christman wants to base his argument on these sorts of facts, why bother appealing first to the agent's judgment about his or her socialization? Why not appeal directly to these objective facts? Furthermore, this would appear to change Christman's procedural account of autonomy into a substantive account, where it is substantive standards of reason that determine what sorts of socialization are acceptable. We have seen in the last chapter the problems with such accounts. So, if the basis for the individual's judgment as to the acceptability of a process of socialization is not based on desire, or on objective facts, then what is left to provide a basis for the judgment? The only option I can think of is arbitrary choice, making a judgment on no basis whatsoever. It seems clear that this cannot possibly be the source of our autonomy, nor does it appear to satisfy Christman's *ab initio* requirement. Thus, it appears that an appeal to beliefs to explain our judgments of our own socialization is a non-starter.

If beliefs on their own cannot license the judgments as to what kinds of socialization to resist, then desires need to be added to the picture. In order for the evaluation of one's own process of socialization to be independent of this socialization it must be based on values and desires that one has independent of socialization, and not formed on the basis of any desires or values formed through socialization. Let us call these desires "socialization-independent desires", or SI desires for short, and the same goes for values.

It seems empirically unlikely that SI desires or values, even if they exist, provide a rich enough foundation for autonomy. We may have various physiologically grounded desires that are largely independent of socialization, such as desires for food and sleep, although it could be argued that the specific character of even these desires is shaped by socialization (for example, what we desire to eat, and when may be shaped by socialization, even if the desire to eat is universal). However, these desires are clearly too rudimentary and limited to provide a basis for a rich and varied autonomous life. The more sophisticated desires we have for particular styles of life and so on seem to be clearly influenced, although not entirely determined by, socialization.

Even if this were not the case, it's not clear that SI desires have any greater claim to being the agent's own than do the ones inculcated through socialization. In order to have a greater claim to legitimacy, SI desires would themselves need to be the result of a process the agent would not resist, but again this decision would need to be based on some set of desires and values, which would again need to be autonomous in order to satisfy the *ab initio* requirement. This problem seems intractable. The only way that an agent could legitimately pass judgment on his or her own socialization is if there were some fact about what that agent's desires and values were in isolation from all causal processes of desire formation. I take it to be highly implausible that such desires are actually possible.

I take this to be a very general problem for any procedural account of autonomy that accepts the *ab initio* criterion. The *ab initio* requirement states that autonomy must be grounded in something that is itself autonomous. This will always invite a regress unless we can identify something whose autonomy is self-certifying. One option is to identify something of this nature external to the subject, such as the demands of rationality, where rationality is given a strongly substantive interpretation. This would be to return to the kind of substantive theory of autonomy we rejected in the last chapter. The other option is to claim that some desires are, or can be, intrinsically autonomous, or that some other non-desire element of an agent's psyche could be intrinsically autonomous and could give rise to desires which would partake of this autonomy. For simplicity's sake I will only discuss desires as candidates for intrinsic autonomy, but I believe that the same comments would apply to anything else within an agent's psyche that was put forward as a candidate for intrinsic autonomy[29]. Now, any desire that is a candidate for intrinsic autonomy will either be caused or uncaused. If it is caused, then it will not be intrinsically autonomous but will instead inherit its autonomy or heteronomy from its cause. It might seem we could get around this by accepting that autonomous desires are caused, but demanding that they be caused by the agent him or herself. However, this merely raises the question of whether they were caused by the agent autonomously or heteronomously, and the regress begins again. This seems to leave only uncaused desires, or an agent outside of causation, as candidate sources for intrinsic autonomy.

There is one other option open to a procedural theory of autonomy that accepts the *ab initio* requirement. We could claim that all autonomous desires are indeed caused, but that these causes circle back and create a loop. Either the desire could be directly self-caused, or else there is a chain of desires causing other desires that at some points loops back on itself. Either of these conditions would amount to a requirement that autonomous agents be self-creating. Then we could legitimately claim that all of an agent's desires are there because of other desires and for no other reason, and thus

[29] Thus, even if my argument against taking beliefs to be enough to form judgments on whether to resist one's socialization are flawed, I think that beliefs run into the same problems with the *ab initio* requirement as desires.

plausibly claim to have identified a type of autonomy that meets the *ab initio* requirement.

The possibility of any such self-creation is metaphysically dubious, however. To give just one reason, for a thing to cause itself to exist at time T, it would seem to need to already exist at time T. But if it already exists, how can it still be in need of being caused to exist? In addition, why should we think that human agents have this unique power to bring themselves into existence? If the *ab initio* requirement demands such a peculiar metaphysical feature, it is unlikely that the requirement can ever be met. It seems that the only option open to us is to abandon the *ab initio* requirement, and accept that autonomy can emerge from non-autonomous elements.

2.3: The Emergence of Autonomy

If we reject the *ab initio* requirement on autonomy, then it is open to us to say that autonomous agents can arise from non-autonomous processes acting on an originally heteronomous agent. I take this to be a perfectly natural conclusion, especially when we consider how children develop into adults. A child below a certain age, it is widely agreed, is heteronomous. The processes of socialization help to shape this original heteronomous individual, and at some point autonomy emerges. One major obstacle to such a view is the problem raised by manipulation cases. For example, Alfred Mele presents the thought experiment of two agents, Beth and Ann, who are both philosophy professors[30]. Ann is autonomous, and is also an extremely industrious philosopher. Beth is also autonomous, but she enjoys a wide range of activities and as a consequence has less time to commit to philosophy than Ann. The Dean of the department makes use of a team of brainwashers to alter Beth's motivational set so that it is identical to Ann's, so that Beth will spend more time and effort on philosophy. We seem to have a strong intuition that such manipulation would render Beth heteronomous, despite the fact that her motivational set has been rendered identical to Ann's, and Ann is by hypothesis autonomous. Since, without the *ab initio* requirement, we accept that Beth could be autonomous even if the source of her desires was not itself autonomous, it seems as if we cannot justify our strong intuition that Beth is heteronomous. This might appear to be a significant cost to rejecting the *ab initio* requirement.

I believe that this worry is partly correct, but that we can assuage the cost of this admission. There are two distinct questions here; the first is whether or not Beth's autonomy was violated, and the second is whether or not after the brainwashing the new Beth is autonomous. It's obvious, in answer to the first question, that Beth's autonomy is violated. If this is not a case of autonomy being violated then nothing is. However, this does not necessarily imply that the new Beth, Beth post-brainwashing, is not *also*

[30] Mele, *Autonomous Agents*

autonomous. Consider what happens with a child in the process of socialization. Many of the techniques commonly used, such as operant conditioning, reliance on beliefs accepted solely on the authority of parents or other authority figures, and so on would be deemed autonomy-interfering if they were applied to adults. However, it doesn't seem overly counter-intuitive to claim that these result in an autonomous agent. Now consider a swamp-man style example; a bolt of lightening strikes a swamp and, by immense coincidence, re-arranges the molecules in the swamp mud into the exact duplicate of a normal individual, right down to the brain structures that contain memories, beliefs, desires and so on. Let's add further that, by coincidence, this swamp-man is physically and mentally identical in every way with another individual who, having grown up and been socialized in the ordinary way, is now autonomous. If we accept that the normally socialized individual is autonomous, why should we not also accept that the swamp version of this individual is also autonomous? I, for one, do not take it to be excessively counter-intuitive to suggest that this swamp man is in fact autonomous, despite the non-standard way in which his autonomy has come about[31].

[31] This conclusion may need to be qualified somewhat. If autonomy is not just a time-slice property, but an agential achievement involving the application of skills over time, as I will argue later, then the swamp man may not have the same degree of autonomy as a regular agent. The swamp man has not had the time to apply the skills of autonomy, unlike the regular agent, and so

To bring the discussion back to Beth and Ann, we should ask ourselves what the differences are between the autonomous swamp man and Beth that trigger our intuitions that the result is heteronomous so strongly. After all, in both cases a sudden, external process produces a psychological system that mirrors another admittedly autonomous one. I can only identify two differences that might be relevant. The first is that Beth's misfortune is the result of the action of a human being, in this case the Dean. The second is that Beth, unlike the swamp man, had a previous mental system that is interrupted. I see no reason that the autonomy of an individual should rest on whether or not the processes that led the individual to have the desires, preferences and so on that he or she has were directed by an individual or were due to natural causes. Furthermore, if human intervention were a barrier to autonomy this would seem to render the results of ordinary socialization heteronomous as well, since socialization is largely the result of human action.

It seems that it is only the fact that, prior to the intervention of the dean, Beth had an existing autonomous mental structure that causes us to judge her heteronomous after the brainwashing. However, I do not think this

might need time in which to do so before counting as fully autonomous. However, the swamp man is not barred from achieving full autonomy due to his bizarre origin. Given time, he has as much chance of becoming fully autonomous as anyone else, and so he still provides a counter-example to the *ab initio* requirement.

alone is enough to call new Beth heteronomous. The old, autonomous mental structure no longer exists, and it is hard to see why the fact that it used to exist should affect our judgment about the new person Beth has become.

To drive this point home, imagine that we have discovered the Dean's nefarious deed, and intend to use our mind-control ray to restore Beth's original mental structure. Prior to doing so, we explain to new Beth what has happened and what we intend to do. Shouldn't we expect her to protest? After all, she is now a copy of Ann, and we would expect Ann to object to being turned into Beth, just as Beth would have protested being turned into Ann. So why shouldn't she claim that she has every bit as much right to protection from mental tampering as old Beth had? It seems plausible that we cannot undo the original wrong without committing an equivalent wrong ourselves.

The mistaken intuition that new Beth is heteronomous has, I believe, two sources. The first is a confusion between manipulation cases in which a person's entire mental structure is re-arranged, as is the case with Beth, which we can call global manipulation cases, and cases where one or more new desires are implanted into an otherwise unaltered mental structure, which we can call local manipulation cases. In local manipulation cases the newly introduced desire is heteronomous, because it will presumably conflict with the existing network of beliefs and desires. Thus, such desires will fail the procedural tests for autonomy. This suggests that in cases of global manipulation the same would be true. However, in cases of global manipulation, the background against which we judged the local manipulation implanted desires heteronomous has disappeared. Thus, if the new network of beliefs and desires implanted meet the procedural requirements for autonomy, the newly created individual is indeed autonomous.[32]

The second source of the misleading intuition is the fact that a terribly immoral act has undoubtedly taken place, and it is clear that this act was the destruction of Beth's autonomy However, I will argue this does not necessarily mean that Beth is not also autonomous after the brainwashing. It is possible for the brainwashing to destroy Beth's autonomy, and yet still leave her autonomous afterwards. To see how this paradoxical claim could be true, we must examine in more detail the effect of the brainwashing on Beth as an agent.

Let us call Beth before the brainwashing Beth$_1$, and Beth after the brainwashing Beth$_2$. Beth$_1$ differs from Beth$_2$ in the entirety of her

[32] The same caveat as mentioned in note 31 concerning the swamp man's autonomy applies here as well, of course. The admission that the post-brainwashing Beth might need time to apply autonomy skills before counting as autonomous, and thus that she will be less autonomous than Ann at least at first, may help to justify our intuitions about the case, so this is an important point. However, the basic point remains that Beth is not blocked from becoming autonomous just because of the bizarre origins of her desires.

motivational structure. Depending on one's theory of personal identity, this may be enough to render Beth$_2$ an entirely new agent. For instance, Derek Parfit's theory where what matters is whether there exist future mental states that are related in the right way, which he calls the R relation, to the agent's current mental states would suggest that Beth$_2$ is not the same agent as Beth$_1$. After all, Beth$_2$'s motivational structure is not similar to nor caused by Beth$_1$'s mental states[33]. Of course, there is more to the R relation than just motivational structure, and Beth$_1$ and Beth$_2$ would still share some R related mental states, such as memories. This might be one of the cases in which there is no determinate answer as to whether Beth$_1$ and Beth$_2$ are the same agent or not.

Similarly, on a narrative conception of personal identity there will be some narrative continuity, in terms of memories and relationships that will survive the change, but also a great deal of narrative disruption. Beth$_1$ will have had many projects and relationships that Beth$_2$ will abandon, due to the new motivational structure that has been imposed on her. Whether or not we judge Beth$_1$ and Beth$_2$ to be different agents, we must admit that a massive disruption has taken place. Any theory of personal identity that takes agency

[33] Derek Parfit, *Reasons and Persons*, (Oxford: Oxford University Press, 1984). Parfit's R relation is not actually a theory of personal identity, which Parfit argues is actually unimportant, but it is a theory of what Parfit takes to be the important kind of agential continuity, and thus it seems appropriate to interpret it as describing the kind of relation between time-slices of agents important for assessing autonomy.

seriously enough to be of interest to a theory of autonomy will have to admit that the disruption in Beth_1's agency is enough to move her into the grey zone of agential continuity with Beth_2, if not render them entirely distinct agents[34].

I don't want to take a stand on whether or not the disruption in agential continuity is sufficient to disrupt personal identity. It is sufficient that this disruption is significant enough to raise such questions, suggesting that we should be prepared to revisit agent-based claims about Beth_1 when applying them to Beth_2. Beth_1 can have her autonomy destroyed, but Beth_2 can still be autonomous, because the two are distinct enough as agents that judgments of autonomy relative to Beth_1 no longer apply to Beth_2. As Noggle points out, autonomy (or authenticity as he calls it) is a two-place predicate. A given desire is autonomous relative to some individual[35], and also relative to a set of mental states. When the background of mental states has changed radically enough, autonomy may be violated without resulting in heteronomy. In this case, Beth_2's desires are heteronomous relative to Beth_1, but autonomous relative to Beth_2.

[34] This will not be true on a biological account of personal identity, but such accounts focus on elements of personal identity that are beside the point for judgments of autonomy. The fact that two agent time-slices are part of the same biologic entity may be metaphysically important, but it has not bearing on the agency of either time-slice, and autonomy is a concept that concerns agency, not biology.

[35] Noggle, "Autonomy and the Paradox of Self-Creation"

Having dealt with the last objection to abandoning the *ab initio* requirement, the question now becomes how the process of socialization can go wrong in such a way as to prevent, or impede, the development of an autonomous individual. One possible answer to this question is that it cannot go wrong in this way; whatever one's upbringing one at some point becomes autonomous. Robert Noggle gives a powerful defense of this view in "Autonomy and the Paradox of Self-Creation." He begins by outlining a rough view of how we might say that some desires were autonomous and others heteronomous. There are some desires that can be easily changed, either by new information or by deliberation. However, we also have a core of more stable desires, which we are slow to change and that serve as a basis for our evaluation of more peripheral desires. Thus, we have a core of stable desires and a periphery of more fluid desires. The core desires can be considered a kind of "deep self."

The desires of the core will change over time, but usually this change will be the result of a natural development directed according to the desires of the core itself. We may, for example, realize that some core desires conflict, in the sense that working to fulfill one tends to frustrate the other, and vice versa. We will then either try to balance the two, or choose to give one up in favor of the other, and this decision will be guided by our core desires. In Noggle's words, these core desires "form the basis and the ultimate court of appeal for the reflective self-adjustment that allows the self to react and develop in response to changing conditions, improved information, and increasing self-awareness."[36] As such, it makes sense to see these changes as internal to the core, a self-development. On the other hand, some kinds of external intervention can disrupt the core desires in a way that is not similarly self-directed, such as head trauma, indoctrination, or mind control rays. These external influences, then, would be heteronomous.

Noggle presents this account as a sketch, rather than a completely worked out theory. The important thing to note, he claims, is that if something like a notion of core desires is taken to be fundamental to autonomy, then prior to the emergence of such a core self it is impossible to violate autonomy. Also, since autonomy is always relative to a core self, it is impossible for such a core to itself be heteronomous. If we combine these two claims, we come to the conclusion that no internalized process of socialization could directly impede autonomy[37]. When a child is undergoing early socialization, there is not yet any core self that could be interfered with, so in terms of autonomy anything goes at this stage. Once the core self is formed, anything that socialization has incorporated into this core self will be

[36] Ibid, 100

[37] Some sorts of socialization might indirectly impede autonomy by, for instance, not encouraging an individual to develop self-control, thus creating an agent who is very weak willed and often acts heteronomously. However, this wouldn't change whether the core desires themselves, whether acted on or not, were autonomous.

autonomous by definition. Thus, if socialization succeeds in implanting oppressive values into an individual's core self, these oppressive values will be autonomous for that individual. Every upbringing will shape a child's desires and beliefs, forming a core over time that will allow for autonomy. No matter how misguided, oppressive, or evil the desires of the core, they will be autonomous for the individual in question. While Noggle admits that *better* selves could have emerged through different socialization processes, the selves that did in fact emerge are no less autonomous just because some other self would have been preferable.

This way of construing autonomy leaves no room for the feminist intuition with which we began. If Noggle's view were right, we would need to abandon the idea that we could criticize oppressive upbringings on the grounds of autonomy. We could admit this point, and fall back on criticizing such upbringings for other faults, for instance the fact that it is oppressive. However, this would leave unaddressed our feeling that the people raised in this way are missing something crucial to autonomy. In order to accommodate both our commitment to the feminist intuition and procedural accounts of autonomy, we will need another way to conceive of procedural autonomy. The next chapter will explore such an alternative conception, one which will allow us to resolve the impasse created by the *ab initio* requirement.

Chapter Three: A Competence Account of Autonomy

3.1: Away from metaphysics, towards competencies

The preceding investigation demonstrates the failure of a particular approach to autonomy. The question of autonomy is the question of what makes someone the author of his or her desires, and so one response is to say that one is the author of one's desire if the desire was not caused by anything external to oneself. This is the kind of reply that the *ab initio* requirement captures. This turns the question of autonomy into a subset of the question of free will[38]. However, if determinism is true, then there are no mental elements that are not the result of external causes. Even if determinism is false, any plausible account of free will must acknowledge that our socialization plays a huge part in shaping our desires, and the procedures of rational reflection that we could use to scrutinize these desires. Thus, even if our choices are not actually determined by our upbringing, they are obviously crucially shaped by that upbringing. This means that the

[38] As pointed out and criticized by Meyers, *Self, Society, and Personal Choice.*

explanation of autonomy will not depend on the answer to the question of free will. We must, in other words, develop a compatibilist concept of autonomy.

Once we abandon the attempt to identify autonomous desires as a kind of self-caused desires, what route is left to address the problem of autonomy? Let us return to one of the paradigmatic cases of heteronomy, the kleptomaniac. Such a person has an intense, perhaps even irresistible, desire to steal, and we judge this desire to be heteronomous. One intuitive explanation for this would be that the kleptomaniac cannot help his or her desire to steal, but this kind of an explanation would lead us back to the *ab initio* requirement we have just rejected. However, another plausible explanation is that the kleptomaniac's desire to steal fails to fit with his or her other goals, plans, and desires. This suggests that autonomy does not depend on the metaphysical status of our desires, but instead on our practical concerns with personal unity and agency. I will refer to this kind of account as an account of practical autonomy, as opposed to metaphysical autonomy.[39]

[39] This distinction is modeled on the distinction drawn between practical and metaphysical theories of personal identity. It is not meant to imply that practical autonomy is "merely" practical, in the sense of being useful rather than true, but just that practical autonomy tracks features of our interest in agency, rather than being based on some natural kind.

Why should we link autonomy with concerns of personal unity? What prevents a more fractured psyche from being autonomous nonetheless? The short answer is that autonomy is a property of agents, and that an individual only counts as an agent when, and to the extent that, the various psychological elements within the individual are unified into a single self[40]. To attribute an action to an agent is to assume that the action expresses the will of the agent as a whole, and is not the result of some deviant fragment. This is suggested by the fact that non-autonomous movements, such as jerks or twitches, are often attributed to the specific body part rather than the agent as a whole.[41]

This might seem to require that the agent's will be considered a further element of the psyche, over and above the various desires and beliefs that it chooses among, but this is not the account I have in mind. Instead, to say that the agent's will is expressed by a certain action just is to recognize that it is produced by a unified mental system, and that it is approved of by

[40] I draw this point from Christine Korsgaard, *Self-Constitution: Agency, Identity, and Integrity*, (Oxford: Oxford University Press, 2009). Korsgaard proceeds to assimilate autonomy to the activity of self-constitution that constructs and maintains agency. I am not committed to this account myself, and in any case Korsgaard follows Kant in grounding morality in this form of autonomy, suggesting that she is interested in what we might call *moral* autonomy, as opposed to the account of *personal* autonomy I am concerned with.

[41] Korsgaard, *Self-Constitution*

that system.[42] In cases of severe fragmentation, there is either no unified mental system to which we can ascribe the actions, in which case there is no agent and thus *a fortiori* no autonomous agent, or else there is more than one such system. If there are several internally unified mental systems, then in so far as they conflict the desires of each system will be heteronomous to the other.

Why must we assume that the two systems will conflict, however? Couldn't we imagine two independent systems that mutually endorse the desires of the other? This might seem like an option, but recall that we only separated the two sets of desires and beliefs into separate systems because they conflicted. If there is not conflict, then there is only one system after all, albeit perhaps imperfectly integrated or expressed.

As a consequence of this change in conception, then it makes sense to consider autonomy a matter of degrees, rather than an all-or-nothing matter. If autonomous desires are picked out by some special property, as in the metaphysical accounts, then it makes sense to insist that all desires either have or lack this property. If instead desires are autonomous due to their contribution to agency, on the other hand, we can expect different desires to satisfy these requirements to various degrees. For instance, the boundary between core and peripheral desires on Noggle's account is bound to be a

[42] I add the caveat to exclude cases where the action is caused by the unified mental system through some deviant causal chain.

fuzzy one, with some desires that straddle the boundary between the two. This suggests that these desires are somewhat a part of the deep self, and so more autonomous than the desires in the periphery but less than those clearly in the core. We can expect a similar view of autonomy as a matter of degrees from all practical accounts of autonomy, which should be kept in mind.

Let us return to Noggle's account of autonomy and see how this basic account can be fleshed out. Noggle speaks of certain desires being central to the individual, and autonomy is assessed in terms of these core desires. However, what makes certain desires core? It cannot be the mere fact that the individual always or usually acts in accordance with a given desire that makes that desire central. Such an account would only give us an account of patterns of behavior. Autonomy, however, is a normative concept, and as such it must be possible to fail to act autonomously. An account in terms of patterns of behavior would fail to account for the possibility of failure, showing only divergence from a pattern. In addition, it would be impossible to account for cases in which a particular desire is intuitively very central to an individual, but is rarely expressed because the situation it deals with occurs only infrequently. It is implausible that my desire for a coffee in the morning, which I almost always act on, is more central than my desire to save the life of someone in distress, which I have never had occasion to act on. I would not be heteronomous in stopping to save someone's life and thus forgoing my morning coffee. Thus, sheer frequency of expression cannot denote the centrality of a desire.

Another possibility would be to explain the centrality of certain desires on the basis of the coherence of the desire with the larger set of desires and beliefs within the individual. Thus, we could say that the desire of the kleptomaniac to steal is heteronomous because this desire conflicts with the larger coherent set of desires that the individual has, such as the desire for social acceptance, the belief that stealing is wrong, the desire to avoid shame and punishment, and so on[43]. However, I do not think that such an account can succeed on its own. In some cases the heteronomous desire will be obviously isolated, as in the kleptomaniac case, but often there will be at least some network of related desires and beliefs on both sides. For example, an individual with a gambling addiction may have a desire to play poker, a desire to risk money, a desire to raise on the flop when dealt a pocket pair, and so on. It seems implausible that the autonomy of this person's desire to gamble will depend on a simple summation of all the desires and beliefs on both sides, especially since it is unclear how we are to individuate desires and beliefs in these cases. Numerical tallying of inter-supporting desires

[43] Ekstrom suggests a view somewhat like this "A Coherence Theory of Autonomy"

seems to miss the fact that a small number of very highly valued desires may be more central to the self than a larger but shallow set of desires.

In order to capture what makes a desire central to a person, we must move away from the purely third person evaluation of desires and consider the first person point of view of the individual in question. It is only from the inside that we can determine what makes a desire central. What makes the kleptomaniac's desire to steal heteronomous is that he or she wishes not to act on the desire, and regrets acting on it after the fact, and so on. In other words, the heteronomy of the desire is based on subjective alienation from that desire, and we can assess which desires are central to an individual on the basis of which desires they identify with, and how strong that identification is, where the strength of identification is a phenomenological matter. Alienation accounts of autonomy are often criticized for being unable to account for manipulation scenarios. However, we have already dealt with that worry above. Therefore, we can identify the "deep self" with those elements of the self that are most strongly identified with, and the periphery is made up of those elements of the self that are weakly identified with, or outright alienated from.

Identification accounts of autonomy (and the related non-alienation accounts of autonomy) are quite common. The general pattern is to attempt to identify some type of mental state from which it is impossible to be alienated from, or to not identify with, and then identify the deep self with an individual's mental states of this type. This approach is visible in the autonomy accounts of Frankfurt, where wholehearted higher order desires play the crucial role; Watson, who substitutes evaluative judgments for wholehearted desires; and Cuypers and Jaworska who both (independently, to my knowledge) use caring as the mental state definitive of autonomy[44]. Let us call all of these accounts, and others that follow this pattern, "intrinsic identification accounts." All intrinsic identification accounts appear open to the objection that we can imagine cases in which people are in fact alienated from a mental state of the type that was supposed to be intrinsically identified with. For instance, in response to Watson's claim that autonomy is constituted by evaluative judgments of first order desires, we can point to the case of Huckleberry Fin, who decides to help the slave Jim escape, despite judging this action to be wrong. Huckleberry acts against his evaluative judgment, but in a way we are inclined to call autonomous. Cases like this lead Cuypers and Jaworska to shift the emphasis from evaluative judgments to caring. Indeed, in the former example Huckleberry seems to act against his

[44] Gary Watson, "Free Agency," *The Journal of Philosophy* 72, no.8 (1975), Frankfurt, "Freedom of the Will and the Concept of a Person" and Harry Frankfurt, "Identification and Wholeheartedness" in *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (Cambridge: Cambridge University Press, 1987), Cuypers, *Self-Identity and Personal Autonomy*, Agnieszka Jaworska, "Caring, Minimal Autonomy, and the Limits of Liberalism" in *Naturalized Bioethics* (Cambridge: Cambridge University Press, 2009)

moral beliefs on the basis of his caring about Jim. However, there are other scenarios in which caring about someone may appear alienated. For instance, we can imagine a case of a woman who feels that she can't help but continue to care about her abusive husband, but experiences this care as coercive and restricting.

I believe that this points to a fundamental error in intrinsic identification accounts. There are no mental states that are necessarily or intrinsically identified with, or non-alienated from. The whole approach is still tied to figuring out some special property of certain mental states that will certify them as autonomous. This approach seems to assume a particular viewpoint on the agent whose autonomy is being investigated, a viewpoint shared by all of the autonomy theories investigated thus far. It is the external viewpoint of an outside observer, laying out an individual's mental states and investigating the intrinsic properties of each in search of what makes some of these mental states autonomous. I wish to suggest an alternative viewpoint that I believe will be more illuminating in formulating a theory of autonomy, which I will call "the deliberative stance." This is the viewpoint of an agent deliberating on what to do. If autonomy is in fact a practical notion, rather than a metaphysical one, then this seems to be the appropriate stance to consider. In addition, the range of obstacles to autonomy will be more apparent from this viewpoint.

The kind of answer suggested by the theories that make autonomy track a particular type of mental state is of no use in deliberative situations, for two reasons. First of all, if we tell someone to act on his or her evaluative judgments, or caring, and so on, it remains possible that there will be a conflict at this level. Someone may find himself or herself drawn in two conflicting directions by caring or evaluative judgments. More centrally, however, it seems to falsify the character of this form of deliberation to reduce the complex self-discovery that accompanies the question "what do I truly want to do?" to the facile answer "act in accord with what you care about," or something similar. We often encounter difficulties in knowing what desire we identify with, and this does not seem to be due to confusion as to which elements of our psyche we inherently identify with. There is something active about our self-discovery that is absent from the passive picture of identification suggested by intrinsic identification accounts, and in general from all accounts that adopt the impersonal third-personal stance, an activity that can only be captured by moving to the deliberative stance.

3.2: Barriers to Autonomy from the Deliberative Stance

A defender of intrinsic identification accounts could raise the following objection. If identification is supposed to be an internal state that picks out an individual's deep self, then why is deliberation necessary? Shouldn't this be transparent to the agent? In other words, how could one fail to recognize one's own deep self? This objection overlooks several complications that complicate self-discovery. One will often encounter cases were two quite deeply held desires or values conflict, and one needs to discover which is more central. Consider the example Sartre gives of a young Frenchman during World War II who must choose between caring for his aged mother and fighting the Nazis[45]. Sartre uses this example to make a point about morality, but we could easily ask instead which action would be autonomous, by asking which of the desires is more central to the young man. It is presumably far from obvious to the young man in this case which decision expresses his true self[46]. However, we need not go along with Sartre in claiming that there is no right or wrong answer, and we must make a radically free choice. Regret, for instance, is a good indicator that we have made a mistake in identifying our deepest desires[47]. Of course, in exceptionally difficult choices like the one in our example, either decision may bring some amount of regret. However, we should distinguish regret

[45] Jean-Paul Sartre, "Existentialism as a Humanism," in *Existentialism from Dostoevsky to Sartre* (New York: Plume, 1956)

[46] We do not need to assume that there will always be a fact of the matter about which option expresses someone's deep self before the choice is made. Sometimes the choice itself will change the person's self, and so beforehand either option would be autonomous. This will be dealt with below when we discuss self-definition

[47] Diana Meyers draws attention to the role of regret in autonomy in *Self, Society, and Personal Choice.*

that we could not do both of two options[48] from regret where we wish that we had made the other choice. One easy way to distinguish these is to ask whether the person would choose differently if placed in the same situation again; if so, then the regret is of the latter kind, and its presence suggests that the choice made conflicted with the agent's deepest desires[49]. Being fully autonomous will require the ability to avoid mistakes of this kind by discovering the relative importance of different desires. Thus, full autonomy requires that we be proficient in self-discovery[50].

The kind of self-discovery in question should not be seen as a passive process in which we try to make ourselves receptive to the relative force of various desires to see which one is strongest. Such receptivity may be one part of the deliberative process, but it is not enough on its own. More active elements of the process would include imagining oneself living with each of the possible decisions, searching for possible sources of bias in one's

[48] This kind of regret as a feature of some ethical decisions is explored by Bernard Williams, "Ethical Consistency," in *Essays in Moral Realism* (New York: Cornell University Press, 1988).

[49] The account becomes more complicated when the agent's deepest desires dictate a course of action that conflicts with the demands of morality. The relation between autonomy and morality falls outside the scope of this thesis, however, so I will assume in all of my examples that none of the relevant options are either morally obligatory or morally forbidden.

[50] Conflict between desires is only one example of how someone might be mistaken about his or her deepest desires. There are many other ways that someone might make such a mistake, such as confusing one's own desires with societal expectations. One example will serve to illustrate the point, however.

decision-making, and so on. My point is not to come up with a definitive list, and indeed the effective strategies may differ between individuals depending on temperament, but merely to draw attention to the active process of self-discovery that is called for by autonomy. This way of understanding self-discovery focuses on the set of skills needed to correctly identify one's own desires. Diana Meyers calls attention to the role of coordinated skills in autonomy in *Self, Society and Personal Choice*. She refers to the systems of coordinated skills as competencies. People are autonomous through the exercise of autonomy competencies. Autonomy is a matter of degrees, just like the exercise of any competency, and the more one develops and applies the relevant skills the more autonomous one becomes. This way of understanding autonomy frees us from searching for a secure bedrock for autonomy, as the intrinsic identification accounts try to do, and can instead account for how different desires and elements of our psyche can be identified with at different times, and how and why we can fail to identify which of our desires are in fact most central to ourselves.

Self-discovery, however, is not enough on its own for autonomy. We must be able to change and develop our deep self, not just identify it. To leave out the element of self-definition in autonomy would be to make autonomy a far too static notion, restricting individuals to a narrow and unchanging authentic self. So an account of autonomy will need to tell us what methods of changing our desires are conducive to autonomy and which interfere with it. However, it should be noted that in attempting to determine what kinds of self-definition accord with autonomy we should not fall back into assuming that self-definition needs to be uncaused or independent of socialization. We can and do define our self on the basis of what socialization has provided. Many accounts of autonomy seem to use a passive model of self-definition, assuming that the individual will define him or herself autonomously by default as long as nothing external intervenes. However, as with self-discovery, I think that this passive model is a mistake. Self-definition is an active process, which requires a set of interrelated skills and makes up a competency, just like self-discovery.

The strength of an individual's desires can change in innumerable ways. These will, by and large, be outside the direct control of the agent. I cannot just choose to change the strength of my desires. However, we are not utterly passive in the formation and development of our desires. We can often predict the results of different actions on our desires, and use this to influence how our deep self develops. For instance, an individual who is anxious in social settings may have no desire to cultivate friendships. However, if this person puts him or herself in social situations often enough, this anxiety will be overcome, and he or she will then desire to cultivate friendships. If the individual is aware that this will happen, then such social acclimatization may be part of a plan of self-development.

Self-definition is important to an account of autonomy because, just as it is impossible to identify certain mental states that are intrinsically identified with, it is impossible to discover a particular process of desire formation or change that is uniquely autonomous. Hypnosis is often identified as a paradigmatic example of heteronomous desire formation, but when hypnosis is chosen as a means of altering unwanted desires is it really that different from the kind of self-habituation discussed in the example above? There are cases where we could imagine hypnosis leading to enhanced autonomy, by eliminating desires that an agent is profoundly alienated from, such as addictive or compulsive desires. The same holds true for all processes of desire formation, under the right circumstances any process could be recruited by an individual in order to promote self-government. This suggests that looking to particular processes to identify features that render them autonomy promoting or interfering is misguided. Instead, we should focus on the skills the agent needs in order to make use of these various desire-formation methods in shaping desires in a way that expresses self-government. We should see self-definition as another set of autonomy skills that complement those of self-discovery, allowing autonomous individuals to manage conflicts of desires, and also develop to adapt his or her desires to changing circumstances in an autonomous way.

There is a third set of skills needed to be fully autonomous, the skills of self-direction. These are the skills needed to put one's autonomous desires into action once they have been identified or defined. More broadly, it also covers the skills needed to identify and direct how one's future conduct will affect one's desires and identifications[51]. These skills will include strategies of self-control to counter weakness of will, the ability to predict and recognize the way that one's actions influence one's character, and the ability to stick with one's plans through at least some degree of adversity. These skills will of course be necessary for someone to act autonomously, and as such are very important for full autonomy. However, they are also important even when we are concerned only with the autonomy of desires themselves. This is because self-direction is essential for developing and implementing the skills of self-discovery and self-definition.

Self-direction is important to self-discovery because some degree of experimentation is necessary to develop the skills of self-discovery. One must be able to act on what one takes to be one's autonomous desires in order to develop these skills. Often, failures of self-discovery will only become

[51] Meyers discusses self-direction in *Self, Society, and Personal Choice*, 59-75. My account is heavily influenced by her discussion, but it does diverge in some places, and the following discussion does not always fit with how Meyers thinks about self-direction.

apparent after one has acted according to what one thought one's desire was. Such emotions as disappointment and regret are important indicators that self-discovery has gone wrong, and such feedback helps to refine the use of self-discovery skills.

Self-direction plays an even more key role in self-definition. The ability to define oneself is, as we saw above, indirect. What an agent chooses to do loops back and has an effect on the agent's desires and identifications. Thus, it is impossible to define oneself if one cannot adopt the courses of action that will produce the desired self. Furthermore, if an agent lacks skills of self-direction then many of the agent's actions will not be the result of the agent's judgments about what he or she most desires. However, these actions will still have an influence on the agent's desires. For example, if, through weakness of will, I fail repeatedly in my efforts to stick to my diet, this may contribute to making me feel that the diet isn't important to me. In this way, failures of self-direction affect my character, and do so in a way that I neither endorse nor control. Thus, failures of self-direction undermine self-definition. Furthermore, the skills of self-direction include the ability to predict ahead of time the effects of various courses of action on one's character. Without such skills it is impossible to employ the skills of self-definition. I may know what I want to define myself to be, but I will not know how to bring this self-definition about. Thus, self-direction is essential to the autonomous shaping of an agent's character.

Together, the skills of self-definition, self-discovery and self-direction constitute autonomy competencies. The competencies approach to autonomy can account for our intuitions on when individuals are autonomous. When we judge someone to be heteronomous, this may be for one of two reasons. On the one hand, the situation described may demonstrate that the individual in question lacks the necessary autonomy competencies. Someone who, when called upon to decide between acting on one of two or more desires frequently chooses impulsively and later regrets the decisions made may be heteronomous due to a lack of self-discovery skills. On the other hand, the situation may not indicate that the individual lacks the necessary autonomy skills, but instead that the situation is such as to interfere with the use of these skills. Thus, a cult leader who uses force of personality, sleep-deprivation, and rote repetition of a given doctrine to create a desire to further the cult's ends renders those so convinced heteronomous by creating situations in which it is difficult or impossible to effectively apply the autonomy skills of self-discovery or self-definition.

If we move from the passive conception of autonomy embodied in autonomy accounts such as Noggle's to the more active, skill based conception suggested by Meyers, then it becomes easier to see how we can substantiate the feminist intuition. On Noggle's account, any self created in childhood is autonomous, no matter how oppressed or evil that self may be. This makes sense if we take a passive view of autonomy. If autonomy is mostly a matter of non-interference with the normal expression and development of the deep self, then autonomy cannot give us any grounds by which to criticize this deep self. However, if autonomy is an active matter requiring competencies to achieve, then not just any deep self will be autonomous. Only a deep self that provides the resources for the development and application of autonomy competencies will allow for autonomy. Self-government, on this account, requires more than non-interference. It also requires the possession of positive abilities. Thus, if we can show that oppressive upbringings undermine the competencies necessary for autonomy, then we will have evidence that those who have internalized the values of such an upbringing are heteronomous. To establish that this is the case, and why, will be the focus of the next chapter.

Chapter Four: Autonomy and Self-Trust

4.1: Autonomy and "Gas-lighting"

The discussion of the competencies approach to autonomy has opened up a new way to consider disruptions of autonomy. Instead of focusing only on the ways that socialization can disrupt a passive sense of identification, we can instead look at how socialization might interfere with the activity of autonomy: the ability to identify and shape one's preferences. I believe that this new perspective gives us the tools to substantiate a version of the feminist intuition. In particular, I think it allows us to develop a better explanation of an example Paul Benson gives of a situation in which autonomy is compromised: a situation he calls "gas-lighting", after the 1944 film *Gaslight* whose plot provides the template for the situation Benson examines[52].

The plot of the movie features a man who marries a young woman in order to find and steal valuable jewels that belonged to her aunt. In order to conceal his intentions he works systematically to reduce his wife to a state of confusion and disorientation. He isolates her from her friends, and engineers situations that encourage her to believe that her memory is faulty, that she is hallucinating, and so on. The character in the film is reduced to a bewildered

[52] Paul Benson, "Free Agency and Self-Worth," *The Journal of Philosophy* 91, no. 12 (1994)

confusion, and strongly triggers our intuition that her autonomy has been compromised.

Benson provides a slightly modified version of this example that clarifies what is at issue. He suggests that the same result could be produced without the kind of malicious intent that characterizes the film. To demonstrate how this might be possible he introduces a new example, which he calls "medical gas-lighting", in which a woman living in the latter decades of the nineteenth century is diagnosed as "hysterical" on the basis of her strong passions and emotional outbursts. This diagnosis, and her subsequent treatment by others as mentally unstable, could be expected to produce the same kind of confusion and uncertainty as in the protagonist in *Gaslight*, and through the same mechanism of convincing her that she was mentally unstable, but we need not assume that anyone in this case was acting in bad faith or attempting to deceive.

Benson argues that procedural accounts of autonomy cannot explain why it is that someone who has been gas-lighted lacks autonomy. After all, neither the protagonist of *Gaslight* nor the medically gas-lighted women appear to have had any interference with their ability to identify with their desires or actions, however such identification is construed. Instead, Benson thinks that the lack of autonomy in these cases is due to the absence of a sense of self-worth. Since this sense of self-worth appears to be a substantive condition on autonomous agency, but is not a particular value commitment and thus seems less constraining than traditional substantive accounts of autonomy, Benson calls self-worth a "weakly substantive" condition on autonomy.

There are, of course, several senses of self-worth that could be at issue here. One sense of self-worth is simply the capacity to match up to some standard against which people may be measured. If I feel I should get excellent marks in school, and I fail to do so, I may feel that I lack a certain self-worth. This is not the type of self-worth Benson is concerned with. Instead, he sees the relevant sense of self-worth as the more Kantian notion of an intrinsic moral status. In order to connect this notion of self-worth as moral status to free agency, Benson cashes out the kind of moral status at issue in terms of considering oneself worthy to answer the normative demands that one believes others would be justified in applying to one's actions. So, for example, the medically gas-lighted women fails to have this kind of self-worth because she believes her insanity renders her incapable of offering intelligible responses to such normative demands, whereas a slave who has internalized the view of his or her enslavers might lack such self-worth because he or she considers him or herself to lack the authority to answer such normative demands due to a fundamental moral inferiority. Lacking this kind of self-worth is taken to undermine autonomy because it leads to the sense that one is unworthy to act. Since someone without this sense of self worth will consider themselves unworthy, and hence presumably unable, to justify their actions to the (perceived) legitimate challenges of others, they will be unable to fully identify with their actions.

Benson's account draws our attention to something important; however, I believe he has erred in identifying the incapacity as a lack of self-worth. Instead, we should understand the barrier faced by those subject to gas-lighting as a lack of self-*trust*. Shifting the focus from self-worth to self-trust will provide us with a more convincing explanation of the particular examples Benson considers, and fits better with the general desiderata of a theory of autonomy. Construing the problem as one of self-trust also allows us to preserve a purely procedural account of autonomy.

The first question, of course, is what is meant by self-trust. Annette Baier has developed probably the most influential account of trust in others.[53] On her account, trusting another to do something is a combination of a belief in the competence of the trusted person and a belief in their good will towards oneself. Good will is necessary because trusting someone, as opposed to relying on them or making use of them, requires situations in which the one trusted has a certain degree of leeway to use his or her own judgment. If the task is specified precisely enough that there is no room for

[53] Annette Baier, "Trust and Antitrust," *Ethics* 96, no.2 (1986)

judgment on the part of the person who is to carry out the task, then the attitude is not one of trust, but merely reliance. Baier identifies this leeway as a condition even of the trust in simple promises that seem not to leave room for interpretation, such as trusting a promise to meet for lunch at noon. This is because one trusts the one who promises to use good judgment to decide when to overrule the promise. If the one individual who made the promise chooses to fulfill this promise even if when some emergency came up that should, by rights, overrule the promise, then this trust is violated, Baier claims. By way of example she says that she

> would feel morally let down if someone who had promised to help me move house arrived announcing, "I had to leave my mother, suddenly taken ill, to look after herself in order to be here, but I couldn't break my promise to you." From such persons I would accept no further promises, since they would have shown themselves untrustworthy in the always crucial respect of judgment and willingness to use their discretion.[54]

This feature of trust seems to present a problem for the notion of self-trust. After all, while it makes sense to have a belief in one's own competence to achieve some task, it is harder to make sense of the notion of believing in

[54] Ibid, 251

one's own goodwill towards oneself[55]. I think the solution is to see Baier's account correctly identifies that all cases of trust must contain both a belief about the competence and the motivation of the one trusted. However, her identification of the correct motivation being in all cases a sense of good will towards the one trusting is better seen as one species of the genus of trust in general.

I think a more general statement of the motivational component of trust is that one must believe that the one trusted possesses good will towards the value one is entrusting to the one trusted. Thus, trusting someone to care for one's children requires a belief that the one trusted is motivated by goodwill not towards oneself, but towards one's child. In this case, the child is the value at issue in the care, but it could be a more general principle as well as a particular object. Of course, you could trust that a friend will care about one's plants, for example, even when he or she has no particular attachment to them because he or she cares about you and you have asked him or her to take care of the plants while you are on vacation. Thus, in some cases the care for the value in question might be derivative of the care for the person who trusts, thus accounting for the wide variety of cases that Baier's formulation of the goodwill component of trust covers.

[55] McLeod raises this difficulty in her account of self-trust in Carolyn McLeod, *Self-Trust and Reproductive Autonomy* (Cambridge: The MIT University Press, 2002)

This seems to provide an account of self-trust in cases of diachronic self-trust. If I am to trust my future self in some particular domain, I must believe that my future self will not only have the competence necessary to live up to this trust, but also that my future self will retain a sense of goodwill towards the ends that I trust myself to achieve. To illustrate this point, consider Parfit's example of the Russian nobleman.[56] The young nobleman cares deeply about reducing inequalities between the nobles and peasants, and so he decides to donate his family lands to the peasants. However, he will not inherit the land until he is older, and he fears that as he ages he will lose his revolutionary zeal and decide to keep the land and be wealthy rather than follow through on his egalitarian plans. Clearly, the young man does not trust his own older self, because he does not trust that his future self will share his good will towards the egalitarian ideals he hopes to foster.

However, there might still seem to be a mystery in understanding cases of synchronic self-trust. It might seem that nothing needs to be explained here; after all, one does not need to trust oneself to do something one is currently doing. I think this is a mistaken impression, however. One way to describe what I'm doing at this moment is typing on a laptop, and I don't need to trust myself to do this, I can just see myself doing it. But another description of my action is that I am writing a master's thesis. Under

[56] Parfit, *Reasons and Persons*, 327-328

this description, I cannot just observe myself to be succeeding at this task (at least not succeeding in writing a good master's thesis). Under this description, even my current activity requires me to trust myself. Any action that is part of a larger teleologically structured activity will call for self-trust on my part in order to reassure myself that my current activity will be efficacious to the achievement of my larger end. This will include a belief that I am competent to achieve this larger end, and that my current actions in fact express good will towards this end.

Someone might object at this point that, while it is perfectly reasonable to doubt the competence of one's current actions to fulfill one's ends, we do not have to worry about our own motivation. We don't need to believe that we are motivated appropriately in our current actions; we can just know that we are, or are not, through introspection. However, I don't think things are so simple. Our own motivations are not always so transparent to us. In many cases we may believe that we are acting for one reason, when it becomes clear in retrospect, or is clear to others, that there is some other motive at play that we fail to recognize. For example, one might think oneself to be a good partner in a relationship, only to realize that, due to insecurity or second thoughts, one has been unconsciously sabotaging the relationship. This would be, on my account, a betrayal of self-trust, the trust one had in oneself to be a good romantic partner.

This account of self-trust covers individual instances of self-trust. Thus, I may have self-trust with respect to some particular skill, so that I trust myself to succeed in its application, or self-trust in terms of my confidence in my ability to resist future temptation, and so on for many other kinds of trust in oneself. Therefore, we must consider what it means to speak of a lack of self-trust being an impediment to autonomy. Clearly it must be lack of self-trust in some central and important areas. No one trusts himself or herself in every circumstance, nor should they. A certain amount of self-distrust is healthy, and plays an important role in self-discovery. A lack of self-trust becomes a risk to autonomy when it is both widespread and resistant to correction.

A lack of self-trust is widespread when it affects several central elements of someone's life. In order to endanger autonomy, the lack of self-trust must be pertinent to many of the day-to-day decisions of the person in question, and must affect decisions that the person takes to be weighty. Day to day decisions are emphasized here because these are the decisions that characterize most of our agency. Hume, sitting in his study, may go through a period of distrusting all of his theoretical conclusions concerning cause and effect, but this does not lead to distrust as to what to do when he leaves his study and goes into the world, and we are not inclined to call him heteronomous. The decisions must be weighty, since I might not trust myself to make good decisions when it comes to buying food, for instance, and this might come up quite often, but if I don't think that making the wrong food-buying decision is that important, then this lack of self-trust will not be liable to undermine my autonomy.

The lack of self-trust must be resistant to correction as well. Someone might go through a period of not trusting him or herself in many central areas of life, for instance when first moving out as an adolescent, but we would not usually describe this as autonomy impairing if the person feels (and is) able to rectify this self-distrust. Such a person still knows what to do; he or she must first act so as to achieve self-trust, and then use this self-trust to tackle the issues that face him or her. Thus, this doesn't seem to impair self-governance. However, in some cases the person will lack self-trust and not feel able to rectify this problem, either because of not knowing what would need to be done to achieve self-trust, or else because the person knows what to do but feels unable to do it. These two conditions, that self-distrust must be widespread and resistant to correction, makes the distinction between healthy self-doubt and the kind of lack of self-trust that can undermine autonomy.

Self-trust as I have described it here is largely trust in one's actions, since it is actions that can be well intentioned and efficacious. However, there is clearly a close link between this kind of volitional self-trust and epistemic self-trust. In particular, a lack of epistemic self-trust is likely to impair volitional self-trust, but not vice-versa. If one does not trust one's epistemic processes, then it will be correspondingly more difficult to trust that one's actions are well intentioned and efficacious, since any belief about such matters will of course be undermined by one's epistemic distrust. On the other hand, I might have every trust in my ability to know what I would need to do in order to trust myself, but I might still lack volitional self-trust, because I doubt my ability to put my knowledge into effect. This could be due to a fear of weakness of will, or else an acknowledgment of a predicted change in one's desires, such as that demonstrated by the Russian nobleman we saw earlier. Nonetheless, most of what I have to say about self-distrust will focus on volitional self-distrust that is mediated by epistemic self-distrust. This is due in part to the fact that the examples in the literature that I wish to explain through self-distrust seem most plausibly to fit this model, and in part due to simple space restrictions.

With this characterization of self-trust in mind, let us examine Benson's gaslight example again. From a purely interpretive standpoint, loss of self-trust seems a more plausible explanation of the gaslight scenario than loss of a sense of self-worth. Consider what the evil husband in the film does in order to induce the protagonist's confusion and uncertainty. He targets and undermines her trust in her memory, by making her believe that she is losing things and forgetting conversations, he targets the reliability of her perceptions, by encouraging her to think she is hallucinating, and he deprives her of the kind of testimonial evidence that might undo the deception, by isolating her from her friends. Perception, memory, and testimony are paradigmatic epistemic capacities, central to our acquisition and retention of knowledge. They are not, however, generally considered central to our moral worth. Thus, there is a strong *prima facie* case to be made that whatever incapacity afflicts the protagonist of the film is of an epistemic character. The medical gas-lighting case is less clearly a case of epistemic tampering, but a strong case can still be made. After all, the catalyst in this case is that the woman is diagnosed with a supposedly serious mental illness. This seems to suggest that she would be reluctant to trust her epistemic conclusions. Indeed, Benson himself seems to support this reading of the problem at times, saying for instance that "she has ceased to trust herself to govern her conduct competently.[57]" This sounds very much like a lack of self-trust, and not a lack of self-worth.

In addition, there are some conceptual reasons to worry about requiring a sense of self-worth in order to count as autonomous. Benson's way of phrasing his account, in terms of one's felt sense of worthiness to respond to others' normative demands, is not as clear as we might like. One

[57] Benson, "Free Agency and Self-Worth," 657

way to be unable to answer another's legitimate moral demands is because one considers one's own action to be wrong. In other words, if I were to decide to shoplift, I would think that in doing this action I would be unable to answer the legitimate normative demands of others, because my action would be normatively indefensible. However, this cannot be what Benson is thinking of. This would, for one thing, render every action that an agent believed to be wrong heteronomous, but surely it is possible to autonomously act contrary to one's moral beliefs. In addition, the gas-lighted woman in his example does not believe that all of her actions are wrong in this way. Therefore, we must conclude that it is not that the person who lacks self-worth feels unable to answer normative demands in the sense that he or she considers him or herself normatively unjustified, or in the wrong. Instead, Benson seems to imply that a lack of self worth means that the person feels unable to give any response whatsoever, unable to show his or her acts to be normatively justified or unjustified. This reading is reinforced by the fact that Benson talks about people being *unworthy*, rather than unable, to answer the legitimate normative demands of others.

If we take seriously Benson's suggestion that the gas-lighted individual feels him or herself unworthy to answer normative demands, this suggests that the person who lacks a sense of self-worth believes him or herself unable to answer normative demands because he or she lacks the moral status necessary to be a legitimate source of answers to normative questions. A person who lacked self-worth in this sense might see his or her situation as similar to that of an inanimate object, which we clearly consider to be unable to give us legitimate responses to normative demands. This view is also problematic, however, because if something is seen as unable to respond to normative demands, it seems to follow that it is an inappropriate target for normative demands. Since Benson's view requires that the person think him or herself unable to respond to *legitimate* normative demands, this means that the person must both think that it is appropriate for others to make normative demands on him or her, but inappropriate for him or her to respond to them, due to a lack of moral status. This seems incoherent; in order to be a legitimate target of normative demands, one must at least be the sort of entity, in other words possess the proper moral status, to respond to these demands.

The most plausible way to interpret what Benson says is that when an individual lacks a sense of self-worth he or she considers him or herself to have the appropriate moral status to be subject to normative demands, but also considers him or herself to lack the competency to respond appropriately to such demands. This seems to fit the case of the gas-lighted woman perfectly; she has the appropriate moral status, but worries that her mental instability renders her unable to respond appropriately. It also fits with the rhetoric often used to support racist or oppressive systems, in which the oppressed group is considered to be stupid or lazy. Each of these characteristics denotes a failure to live up to some normative requirement, implying that members of some group are incompetent to live up to their normative duties on their own, and therefore need to be guided or controlled by others. However, this reading of self-worth seems to collapse into the self-trust requirement suggested earlier. Presumably, a lack of competency to live up to normative requirements is a failing of theoretical or practical reason, and so someone who feels that they lack this competency will lack trust in his or her theoretical and practical deliberation. Thus, insofar as Benson's requirement is plausible it is more naturally labeled a kind of self-trust.

Even if the preceding critique of Benson's self-worth criterion is not taken as definitive, I think we have another reason to prefer an explanation in terms of self-trust. The self-worth criterion is offered as a substantive requirement on autonomy. According to the self-worth criterion, in order for an individual to count as autonomous his or her mental system must include a sense of self-worth, and thus autonomy is not explicable in purely procedural terms. The addition of a substantive condition on autonomy is introduced in order to explain our intuition in the case of the gas-lighted woman and other similar situations. However, I take it that our confidence in the intuition that the people in these scenarios lack autonomy is stronger than, and in large part the basis for, our confidence in Benson's particular account of the source of this heteronomy. Thus, if we could explain the gaslight cases in some other way, Benson's account would lose much of its motivation.

As we saw in the first chapter, substantive accounts of autonomy come with substantial theoretical costs. If accepting these costs were the only way to explain our intuitions in the gaslight cases then this cost might have to be accepted. However, if self-trust can justify our intuition in the gaslight case while preserving a purely procedural account of autonomy, then this account will allow us to preserve more of our theoretical commitments, and will therefore be more convincing from the standpoint of reflective equilibrium. Thus, even if we take Benson's self-worth criterion as a possible explanation of the gaslight case, we still have strong reason to prefer the self-trust account, since it satisfies the concerns that motivate Benson's account without incurring the theoretical costs.

4.2: Lack of Self-Trust as a Barrier to Autonomy

Since the case of gas-lighting points us towards self-trust as a condition on autonomy, we must now investigate why a lack of self-trust should impede autonomy, and why this impediment should be seen as a procedural barrier. According to the competencies approach to autonomy discussed in the last chapter, being autonomous depends on the active application of a set of coordinated skills of self-discovery, self-definition, and self-direction. Now, there is nothing conceptually impossible about an agent who consistently and successfully applies all of these skills despite lacking all self-trust. Such an agent is fantastically unlikely, however, given what we know about human psychology[58].

Let us start by seeing why a lack of self-trust would disrupt the skills of self-discovery. First of all, a lack of trust in a conclusion tends, in most circumstances, to undercut the stability of the judgment. Thus, even if an agent who lacked self-trust did successfully apply autonomy skills to determine which of a set of conflicting desires[59] was autonomous for him or her, this conclusion would tend to be unstable, since it was not taken to be trustworthy. Particularly worrying from the standpoint of those oppressed is that the correct judgment about what is an autonomous desire for the agent

[58] My account of how a lack of self-trust is a procedural barrier to autonomy resembles and builds on Trudy Govier's account in Trudy Govier, "Self-Trust, Autonomy, and Self-Esteem" *Hypatia* 8, no.1 (1994).

[59] I take it that two desires conflict when the satisfaction of one tends to frustrate the satisfaction of the other and vice versa. This can either be because the satisfaction of one is actually inimical to the other, such as the desire to shoplift and the desire to be a law abiding citizen, or more contingently because the satisfaction of one desire will predictably consume resources that will rule out the satisfaction of the other, such as when one must decide whether to go to a movie or read a book in the evening, where one only has enough free time to do one and not both.

might be replaced by what others tell the agent, falsely, is his or her autonomous desire. Thus, an agent without self-trust will be vulnerable to having their autonomous desires undermined and replaced by social convention or the pressure of those around them, and thus rendered heteronomous.

Another way that a lack of self-trust can undermine the application of autonomy skills is through eroding the motivational force of successful applications of autonomy skills. One of the functions of autonomy skills is to allow people to resist the temptation of desires that are not autonomous for them. The knowledge that a given desire is more autonomous than a competing desire can bolster our motivation to act in accord with the autonomous desire. Consider a common sense example; I may find myself on a diet, but tempted nonetheless to eat a slice of cake. Indeed, my motivation to eat the cake might be significantly stronger than my (current) motivation to stick to my diet. However, I may bolster my resolve by reflecting on which desire is more autonomous, or more reflective of my "deep self". Resolving that my desire to stick to my diet is more autonomous, I may find that such a consideration bolsters my motivation and allows me to make my desire to stick to my diet efficacious. However, if I do not trust my judgment that sticking to my diet is really what I want to do, then it seems to be a psychological fact that I will find that such considerations are motivationally useless, or at least less useful than otherwise, and thus I may end up eating the cake anyway. This represents another way that heteronomy can be produced by lack of self-trust.

Both of these considerations apply to someone who lacks self-trust but manages nonetheless to apply autonomy skills successfully. Both cases consider how someone might successfully determine what desire is autonomous for him or herself, and then proceed to consider how this success might be compromised after the fact, by being replaced by heteronomous desires in the first case or by being motivationally inert in the second case. However, the largest obstacle to the autonomy of those who lack self-trust is the fact that such people are unlikely to develop or apply autonomy skills. Why bother spending the effort to imagine various options, see how they resonate with one's life plans, and so on if one views the result as little more reliable than a blind guess? A lack of self-trust will often lead an agent to forego any attempt to apply autonomy skills, allowing the course of his or her life to be dictated by another or by reactions to situations as they arise, without trying to integrate these into a larger pattern. Obviously, this is a matter of degrees. Very few agents have self-trust undermined to the extent seen in the gaslight case. In less severe cases we can expect the agent to apply autonomy skills to some extent, but less fully and consistently than if self-trust were present.

The same kinds of problems that plague a self-distrusting agent's attempt to be synchronically autonomous also recur at the diachronic level. Recall that on the competency account of autonomy, it is not just one's current psychological set that determines what is autonomous for the individual, but also the ongoing activity of deliberation and decision that changes and develops the self in accord with autonomy, through the skills of self-definition. A lack of self-trust endangers both the development and the application of such skills. If an individual lacks trust in the output of his or her deliberative process, then the kind of deliberation that goes into self-definition will appear useless. After all, why bother spending time and effort to determine which sort of life will most truly suit you if you do not take yourself to be a competent judge of the matter? Thus, a lack of self-trust will tend to encourage a more passive attitude towards one's future, and such a person will tend not to develop the skills necessary for self-definition. In addition, to the extent that the individual does develop and use these skills, they will be unlikely to stick to their conclusions in the face of hardship or opposition from others. Even if they manage to be autonomous, they are likely to be dislodged back into heteronomy when challenged. These difficulties mirror those that we saw in the application of self-discovery skills when self-trust is lacking.

It is worthwhile to note here that none of the preceding discussion hinges on whether the lack of self-trust at issue is warranted or not. In the gaslight case, the tragedy of the protagonist's situation is highlighted by the fact that her distrust is unwarranted; we as the audience know that her judgments are fully trustworthy. However, even if she had in fact been going insane, the effect of realizing this would have the same undermining effect on her autonomy. This complicates matters when we consider the medical gas-lighting case; if we assume that the doctor in the example is acting on good faith, then it seems that no one has actually acted wrongly. It seems unavoidable, given our limited knowledge, that we will sometimes consider people unworthy of self-trust mistakenly, and thus might contribute to undermining their autonomy wrongly. Perhaps all we can take away from this is that we should acknowledge the autonomy-undermining features of self-distrust and thus err on the side of caution in such cases. We might need to set a higher standard of evidence before acting in ways that undermine self-trust, in recognition of the grave moral consequences of being mistaken in these cases.

Of course, we should be careful to recognize the lack of self-trust, and the resulting impediment to autonomy, is a matter of degrees. The gas-lighted woman is a particularly egregious example of the lack of self-trust; it is likely that few people suffer such a complete erosion of confidence in their own judgments. However, the greater the extent that an individual lacks self-trust, the more likely that the problems identified earlier will sometimes result in local instances of (greater degrees of) heteronomy, and the less autonomous the person will be as a whole. This conclusion is also entirely consistent with procedural accounts of autonomy. A procedural account of autonomy, recall, is any account that makes autonomy depend on the formal features of the agent's mental structure. The competencies approach to autonomy is procedural, since what makes certain preferences autonomous is that they have been discovered and/or created by the application of the autonomy skills of self-discovery and self-definition. Since these skills themselves are formal, a set of tools that can be applied to any preferences and that do not prejudge the outcome, competency approaches to autonomy are procedural. However, as the investigation of this chapter indicates, the successful operation of autonomy skills requires that one possess some level of trust in oneself. While this does create a non-formal requirement, this is due to the practical necessity of such a belief to meeting the procedural constraints on autonomy. It might be argued that adding such a non-formal requirement makes this account substantive after all. However, I wish to resist this conclusion. The relation between self-trust and autonomy is contingent, not conceptual. It is perfectly conceivable that there would be some sort of being that could be autonomous without any self-trust. There might turn out to be other preferences or beliefs that autonomy always contingently requires for human beings. For instance, it might turn out when self-discovery skills are applied diligently by human beings it always turns out that a desire for independence is discovered. If this were so, then it would turn out that possessing a desire for independence is necessary to be fully autonomous, but this would not be a substantive condition on autonomy. The crucial point is that the procedural account is doing the explanatory work, without reference to any substantive requirements in framing the theory. The fact that some substantive requirements do in fact fall out of the theory as *conclusions* does not change the procedural nature of the theory. Thus, we can preserve the advantages of procedural accounts while doing justice to Benson's observations about the gaslight case and similar examples.

While I think that Benson's examples indicate the need for self-trust rather than self-worth, I do not think it impossible that some notion of self-worth might play a similar *procedural* role in permitting the development and application of autonomy skills. The same goes for the claim that self-respect is necessary for autonomy, as suggested by Robin Dillon.[60] More generally, it seems likely that a variety of self-regarding pro-attitudes may be procedurally necessary for autonomy, although the details of the pro-attitude

[60] Robin Dillon, "Towards a Feminist Conception of Self-Respect," *Hypatia* 7, no.1 (1992) and Robin Dillon, "Self-Respect: Moral, Emotional, Political," *Ethics* 107, no.2 (1997)

and the role it plays in the deployment of autonomy skills would need to be spelled out in each case, a task that is far too extensive for this thesis.

It is interesting to note that self-worth, self-respect, and self-trust have all been interpreted as *substantive* conditions on autonomy[61]. Since it is impractical to discuss the merits of each self-referential pro-attitude as a substantive condition on autonomy individually, let me make some general comments as to why I suspect that a procedural analysis of each of these attitudes will be superior. It appears to me that in each case in which a self-referential pro-attitude is offered as a substantive condition on autonomy, there are two primary motivations. The first is to explain why certain types of lives that are characterized by subservience or other disliked features, and which manifestly lack the pro-attitude in question, cannot be autonomous. This motivation is misplaced to my mind, and motivated by a failure to keep separate the notion of autonomy as a characteristic of agents and the quite different conception of autonomy as an ideal to strive towards. There are numerous resources we can draw on to criticize certain types of lives as less valuable or choice worthy, not all such criticism need be in the language of autonomy.

[61] Self-worth, as we saw above, by Benson, self-respect and self-trust by Carolyn McLeod in *Self-Trust and Reproductive Autonomy* (Dillon herself does not use the procedural/substantive autonomy in her discussion of self-respect, so it is difficult to tell how she would characterize the relation of self-respect to autonomy in her work).

The second primary motivation comes from observing some impairment of resolute agency in those who lack the proposed pro-attitude. People who lack the relevant attitude may seem indecisive, irresolute, or frequently regretful of their choices. These kinds of failures of agency trigger our suspicion that these people lack autonomy. However, these kinds of practical failures of agency strongly suggest that some procedural failure to apply autonomy skills is what is really at work. After all, it is not the mere absence of the relevant pro-attitude that is responsible for our judging the individual heteronomous, but rather the failures of agency that this absence results in. This suggests to me that procedural accounts of the influence of self referential pro-attitudes on our agency will better explain why those who lack these pro-attitudes fail to be autonomous.

We can use the dependence of autonomy on self-trust to account for the feminist intuition we began with. It is the effect of oppression on self-trust that accounts for the heteronomy of those who internalize oppressive norms. The next section will demonstrate how this can occur, and why it is a convincing construal of the examples of heteronomy of those oppressed.

4.3: Heteronomy, Oppression, and Epistemic Injustice

In order to account for the feminist intuition by appealing to a lack of self-trust, we need to identify a mechanism that would generate such a lack of self-trust in all those raised under the various disparate conditions that are grouped together under the label oppression. One tempting reply would be to just cite the fact that much oppressive socialization explicitly includes a belief in the untrustworthiness of certain individuals as one of the elements of the belief structure inculcated. For instance, women in western societies were, and in some cases still are, taught to think of themselves as overly emotional and lacking objectivity, and combined with the cultural denigration of emotion and valuation of objectivity this seems to directly implicate a belief in the untrustworthiness of women. To the extent that women internalize this belief, they will lack autonomy.

This seems like a satisfactory answer in some contexts, but it also seems too narrow. The feminist intuition seems to stretch to many cases in which there is no such explicitly inculcated belief in the untrustworthiness of the oppressed individual. In western society today the kinds of prejudice that contribute to oppression rarely take the form of explicitly inculcated beliefs in the inferiority of any group of people. Women are not in general told that they are less competent than men; instead, the prejudice against women's competence is implicit in differential treatment. The same goes for other oppressed groups, by and large. Thus, we need an explanation for how these implicit prejudices undermine self-trust. I believe that such an explanation can be found in Miranda Fricker's idea of epistemic injustice[62]. It is the consistent subjection to epistemic injustice that is responsible for the lack of self-trust of the oppressed, and thus for their lack of autonomy.

Fricker characterizes epistemic injustice as a particular kind of injustice in which "someone is *wronged specifically in her capacity as a knower*."[63] She specifically discusses two kinds of epistemic injustice; testimonial injustice and hermeneutic injustice. Either variety can undermine self-trust, and thus autonomy. Testimonial injustice is characterized as a prejudicial dysfunction in the credibility attributed to a speaker. In other words, the speaker's testimony is taken to be less likely to be true than the evidence warrants, and this undervaluation is due to prejudice rather than innocent error. Of course, this will not always be enough to undermine acceptance of the offered testimony. If the offered testimony is highly plausible, then it may still possess enough credibility for belief even given the unjustifiably lowered credibility assessment. However, the victim of the epistemic injustice has been wronged even if the testimony is believed, for it is believed with less assurance than should have been the case.

[62] Miranda Fricker, *Epistemic Injustice*, (Oxford: Oxford Univeristy Press, 2009)
[63] Ibis, 20

The central case of testimonial injustice is what Fricker calls systematic identity-prejudicial credibility deficit. Identity-prejudicial credibility deficit refers to cases in which an individual's identity as such causes hearers to unfairly deflate their assessment of the speaker's credibility. Such testimonial injustice is systematic if it tracks the individual across various social contexts. While there may be instances in which a localized deflation of credibility is quite damaging, or where a credibility excess actually functions as a harm, these will be the exception rather than the norm[64]. Generally, systematic identity-prejudicial credibility deficits will be the most harmful cases of epistemic injustice. In addition, it will become clear that it is the central cases that are most explanatory for situations of oppression. Thus, from here on in when I use the term testimonial injustice I will be referring to cases of systematic identity-prejudicial credibility deficit.

In order to establish that credibility deficits of the kind just discussed are a genuine injustice we must also say what specific harm they do to those subject to such injustice. Of course, there may be various disadvantages to

[64] It might seem that every credibility deficit, at least if it tracks a group, results in a credibility excess to some other group. This isn't necessarily the case, however, for credibility attribution is not a zero sum game. Every statement has some quantity of credibility due it, and I can give some people less than they are due without giving more credibility to anyone else than they are due. Of course, this will function as a credibility excess if it is a case of pitting the word of someone whose credibility I unfairly deflate against someone who I give proper credibility to, but not all cases of credibility deficit feature such weighing of competing testimony.

not having one's ideas taken seriously, such as difficulties with career advancement and so on, but these are properly seen as extrinsic to the *epistemic* injustice; they are further injustices that stem from this. The more purely epistemic injustice is the harm that these testimonial injustices do to the knower's confidence in herself as a credible testifier. Fricker offers the example of a philosophy professor of Mexican descent who was the target of persistent ungrounded complaints from a white graduate student, and who received no support or encouragement from her colleagues until a white senior professor had the same problem with the grad student. In the process, the woman suffered from severe self-doubt due to the general dismissal of her credibility[65]. Assuming that lack of credibility afforded her was due to her identity, this is a clear injustice. It is also a clear example of how testimonial injustice can erode epistemic self-trust, which as we have seen will entail a loss of volitional self-trust as well.

This kind of testimonial injustice in the course of early childhood socialization can be even more devastating. As an adult, one has some settled view as to one's credibility that can be used to resist, to some extent, the undermining effect of testimonial injustice. As a child, however, one is still discovering the social world, and the accepted rules whereby one may contribute to it. Therefore, if one is subjected to repeated testimonial

[65] Ibid, 48

injustices as one grows up, it is very easy to conclude that one's statements do indeed warrant the lower credibility bestowed on them by others, and thus to lose self-trust in one's judgments.

This account of testimonial injustice seems most plausible in the cases where the credibility deficit is justified on the grounds of the lesser competence of those with the identity that is prejudiced. However, as Fricker points out, testimonial injustice may be based instead on a distrust of the motives of the individual whose credibility is devalued. An example of this type of prejudice is the type of prejudices that have been used against Jews. Jews are not displayed as less intellectually competent, but instead as liable to various moral deficiencies, such as greed and a manipulative nature. It is less clear at first how someone subjected to this type of testimonial injustice could have their self-trust undermined. After all, it is not their competence that is in question, so why would they come to distrust the conclusions they come to?

Self-trust in these cases is undermined nonetheless, because such a person may come to distrust their ability to accurately discern their own motives. As we have seen, self-trust requires a belief about one's competence, but also a belief about one's goodwill. To the extent that such a person internalizes the prejudices that surround him or her, he or she will come to distrust his or her judgment about motives, and thus lack the belief in one's own goodwill towards one's goals. After all, on introspection such a person will think that they have one motive, but will find that those around him or her consistently assign him or her a different, less laudable motive. If the individual internalizes the prejudice of the wider society, then he or she will believe that his or her motives are in fact the ones attributed to him or her by the wider society, but since this clashes with the motives revealed by introspection, the only conclusion such a person can draw is that he or she is unable to reliably discern his or her motives.

This lack of trust in one's motives is equally detrimental to autonomy as the distrust based in a belief in one's incompetence. While such a person may have trust in the fact that the output of his or her deliberation will achieve his or her ends, he or she will not be confident of what these ends actually are. Indeed, such an individual will be inclined to think the ends are malicious. Thus, a distrust in one's motives is as effective in undermining self-trust as a distrust of one's competence, and has the same results.

Fricker also discusses another sort of epistemic injustice, hermeneutical injustice, which can also undermine self-trust. This kind of injustice occurs when the hermeneutical resources of their social context unfairly disadvantage some individuals. The hermeneutical resources available to a person disadvantage him or her when there are gaps in these hermeneutical resources that prevent him or her from understanding or articulating important features of his or her experience. This disadvantage will qualify as an injustice if the lack is not coincidental, but is instead sustained by the interest some elements of the society have in rendering these areas of social experience inarticulate.

The example Fricker gives is of the hermeneutical gap that existed prior to the development of the concept of "sexual harassment." She discusses the plight of Carmita Wood, who worked at Cornell's department of nuclear physics, but quit due to the persistent sexual harassment of a senior professor. When applying for unemployment insurance, Wood was challenged to give a reason for leaving her previous employment. She found it impossible to articulate her experience, and ended up just saying she left for personal reasons. Her unemployment request was subsequently denied. Here we can see both types of harms that hermeneutical gaps can create. From the information we are given, Woods seems unable to articulate or discuss her own situation. While she feels strongly (and accurately) that what the professor subjected her to is wrong, without a socially accepted concept within which the wrong fits it is difficult for her to articulate the wrong to herself or others.

Equally harmfully, the lack of the hermeneutical resources in the wider culture deprives Wood of access to the normal societal means of redressing wrongs. Since what happened to her does not fit into the existing categories of wrongs, she has little access to the social resources that might help her. This is demonstrated by the fact that she is turned down for unemployment insurance; her reasons for leaving don't fit any of the recognized reasons, so she is denied resources that would help her recover from the wrong done to her.

Wood's story has a happy ending; her case was one of the catalysts of the feminist movement that established sexual harassment as a recognized category of wrongdoing in our society, and so helped correct the hermeneutical gap that had disadvantaged her. However, her case provides an excellent example of how hermeneutical gaps can harm people whose social experience is thereby rendered inarticulate and unrecognized. Of course, as Fricker notes, there is a sense in which both the harasser and the harassee are equally cognitively disabled by the hermeneutical gap; but only the harassee is disadvantaged by this gap. This asymmetry can be accounted for by noting that the harasser's goal is actually furthered by the hermeneutical gap, but for the victim the gap is disabling and harmful for the pursuit of his or her goals.[66]

Intermittent hermeneutical gaps are inevitable. As circumstances change, new social experiences will become possible, and some of these will fall outside the hermeneutical net that exists at a given time. This only

[66] Ibid

becomes an injustice, and also has a serious impact on self-trust, when these hermeneutical gaps are systematic and widespread in an individual's social experience, and these gaps interfere with important interests. This is most likely to occur when the individual hermeneutical gaps are part of a larger system of hermeneutical marginalization. Fricker defines hermeneutical marginalization as "when there is unequal hermeneutical participation with respect to some significant area(s) of social experience."[67] The powerful and socially advantaged have disproportionate access to the means whereby hermeneutic understandings are generated, and this can lead to systematic hermeneutic gaps in areas of social experience that are experienced mainly or exclusively by those without such resources. Furthermore, in respect to some areas of experience, it is in the interest of those in power to prevent any clear understanding from emerging. For example, the idea that repeated sexual advances in the workplace are always instances of flirting, and thus harmless, has a positive value for some powerful individuals, and they may be motivated to work to undermine any improvement in the hermeneutical situation[68].

Hermeneutic injustices due to widespread hermeneutic marginalization are another means whereby self-trust can be compromised. The effect is more indirect than in cases of testimonial injustice. In

[67] Ibid, 153
[68] Ibid, 152

testimonial injustice, the person is directly confronted with others devaluing his or her epistemic competence. In cases of hermeneutic injustice, on the other hand, the individual finds him or herself with significant areas of social experience that are inarticulate and likely to be dismissed by others. To such a person, the lived experience of the social world will consistently diverge from the concepts and frameworks available to describe this world. Thus, whenever a person subject to hermeneutic injustice attempts to describe their own social experience they are likely to either be stymied by the lack of vocabulary, or dismissed as incoherent or confused. In some cases such a dissonance may inspire people to reform the hermeneutical situation they find themselves in, as in the case of Carmita Wood who, in collaboration with like-minded feminists, managed to have sexual harassment recognized as a legitimate hermeneutical category. However, most people will lack either the opportunity or the support system to do this. If the existing hermeneutical framework is internalized and accepted, then the lived experience must be distrusted.

What does it mean to distrust one's lived experience? In our ordinary lives, we experience any number of affective cues that help orient us to our situation and provide a basis on which to discover what we desire. For example, I might think that I desire to study physics in university, but I notice that whenever I set myself to studying a physics textbook I become restless and irritable, while when I read philosophy I am able to focus for long periods of time, and find myself thinking about what I have read for a long time afterwards. This is a clue that perhaps pursuing physics is a mistake. These affective cues are not yet themselves desires, for they lack propositional or intentional content. However, they provide important data in the process of determining what is autonomous for an individual. However, not all such cues are reliable. Sometimes affective cues need to be ignored or worked around, and in such cases these cues are treated as untrustworthy data. We can now see the harm of hermeneutical injustice to autonomy. Wood feels a sense of discomfort and unhappiness that suggests (correctly) that she is being wronged; were she to internalize the insufficient hermeneutic resources of her society she would have to conclude that these affective cues are untrustworthy and should be ignored. If it persists, which it presumably would, she must judge herself to that extent irrational, in the same way as someone suffering from paranoia whose affective cues of danger would be systematically irrational. This same pattern would be repeated in other areas of her life affected by hermeneutical injustice. If these areas are suitably widespread, then this entails a similarly widespread irrationality, and untrustworthiness of one's affective cues in general. However, as we have seen in the discussion of the role of regret in self-discovery in chapter two, deliberating in such a way as to secure autonomy requires a careful attention to affective cues that would signal which desires are more representative of one's true self. Distrust in this area would thus undermine the ability to rely on such cues, and the kind of self-distrust linked to hermeneutic injustice would be disruptive of autonomy skills, just as much as the distrust created by testimonial injustice.

Does a lack of self-trust caused by epistemic injustice actually justify the feminist intuition? I am inclined to think that it does, at least in the instances we want most to account for. Let us briefly return to Stoljar's example that motivated the feminist intuition, the example of women who take contraceptive risks. Stoljar discusses how the women who took contraceptive risks engaged in a process of tacit bargaining, weighing the costs, both social and personal, of violating norms of female sexuality and femininity versus the costs of taking a contraceptive risk. Now, such bargaining in itself is clearly not a barrier to autonomy, indeed when the norms one accepts conflict it is likely to be a valuable autonomy preserving strategy. However, Stoljar objects that when the norms internalized are false and oppressive, as are the norms of female sexuality that proscribe women from using effective contraception because it is wrong for women to plan for sex, then such a bargaining process is in fact heteronomous.

If we take a procedural conception of autonomy, as I have argued for, then we will have to abandon this global claim and accept that it is possible for an agent who accepts false and oppressive norms to nonetheless be autonomous. Unless we define oppression in such a way as to entail a lack of self-trust, it will be possible for someone to accept norms that are to some degree oppressive without losing autonomy. For example, a woman might accept that her role is to raise children, perhaps based on religious beliefs, and thus choose to privilege this role even when she feels no attachment to it, and it conflicts with other strong desires she has and leaves her feeling less fulfilled in her life. It seems that such a woman accepts a norm that is oppressive to her. However, there is nothing in this story to rule out the possibility that she possesses all of the autonomy competencies we discussed in chapter two, and has applied them successfully. Her beliefs about the role of women may be deeply identified with and central to her identity. Even if her values are based on false beliefs, and despite the fact that she would be happier were these false beliefs to be corrected, there is not an indication that this woman is acting contrary to her deep self.

However, I think that the women in Stoljar's example may well turn out to be heteronomous, although more information would be needed to be sure. Unlike the hypothetical autonomous oppressed woman, who is resolute and confident in her decision to live by her oppressive values, the women interviewed by Luker display signs of confusion, vacillation, and indecision. These are not the traits of an autonomous agent, and this raises suspicion that these women are not autonomous. In order to substantiate this suspicion we need to look closer at the details of what the women interviewed say about their own decisions, and compare this to what we might expect to see if the women in question lacked self-trust.

If the women in Luker's study lacked self-trust, they would feel less than fully competent to deliberate on and decide between the norms that they had accepted. They might choose to observe one set of norms when alone, such as following the norm proscribing sex before marriage by not taking contraceptives, but be easily pressured into following a conflicting norm by others, such as the pressure of a romantic partner to prove commitment or femininity by having sex. This would be an example of the way that a lack of self-trust leaves one without confidence in one's own decisions, making them vulnerable to the pressures of social expectations or the pressure of specific others. Alternatively, women lacking self-trust might attempt to fulfill the demands of all of the conflicting norms by adopting a compromise position, such as having sex only "spontaneously", thus trying to hold a middle ground between religious values of chastity and social norms of sexual availability. This response could be explained by the lack of development of autonomy skills due to a lack of self-trust. It could result from inadequately developed self-direction skills that prevent the women from finding more effective ways to balance conflicting obligations without compromising autonomy, or from a lack of self-definition skills that prevent the women in question from shaping a more consistent set of values, or from a lack of self-discovery skills that would reveal one or more of the values to be mere societal pressure without a basis in the women's deep self. These two strategies make sense if the women who take contraceptive risks lack self-trust, and they seem to match the stories the women tell in Luker's book. This is not to deny that a compromise between conflicting values could be a rational and autonomous choice, only to point out that it could equally be motivated by a lack of self-trust that precludes a more decisive commitment to a single set of norms.

Whether the women in question actually lack self-trust is of course an empirical question, and not one I am in a position to speak on in the abstract. However, it seems that in general cases in which a course of action is adopted confidently and resolutely will not generate the feminist intuition. It is only when agency seems impaired by irresolution and vacillation that we are apt to become suspicious of the autonomy of the action. Such vacillation fits the account of a lack of self-trust, and this suggests that most instances of the feminist intuition can be accounted for by a lack of self-trust.

4.4: Conclusion

The cases identified by Stoljar and Benson have brought to light an important element of autonomy. The real life example of women who take contraceptive risks and the fictional case of the gas-lighted woman show us agents who seemed to have less than full self-governance, but who don't seem to be missing any of the formal elements identified by most procedural accounts as necessary for autonomy. In both cases, this is interpreted as requiring the addition of some kind of substantive condition on autonomy. However, what both cases actually point us towards is the need to move to a more dynamic conception of the requirements of autonomy, such as the competencies approach developed by Diana Meyers, and the recognition of the importance of self-trust to achieving autonomy on this account.

The importance of self-trust for autonomy also allows us to account for the connection between oppression and heteronomy. In the end, I believe that we must abandon the view that the link is conceptual or necessary. It is not impossible to internalize oppressive values and norms and still be autonomous. However, recognizing the importance of self-trust allows us to explain the contingent connection that justifies the intuition that many people who have internalized such values are heteronomous. This is because the internalization of such values tends to either entail or at least accompany a lack of self-trust. Most oppressive systems subject the oppressed to widespread epistemic injustices, which will in most cases undermine self-trust.

This conclusion has both theoretical and practical implications. In theoretical terms, it helps to account for what Stoljar calls the feminist intuition. Since this intuition is deployed to discredit procedural accounts of autonomy, accommodating it under a procedural theory helps reinforce the case for sticking to procedural accounts of autonomy, especially given the problems with substantive accounts that I discussed in chapter one.

Practically, the implications are more complicated. While those who lack self-trust will be less autonomous, it is unclear how best to deal with this situation. If this is taken as a reason to overrule their decisions in medical, legal, or other similar contexts then this is likely to only reinforce the lack of self-trust that created the problem, by signaling that such people are not competent to make decisions for themselves. Determining the precise means whereby the existing minimal autonomy skills in those who lack self-trust could be nurtured will be a complex task, and well outside the scope of this work. Nonetheless, hopefully recognizing a lack of self-trust as a barrier to autonomy will help clarify the difficulties involved in identifying and overcoming heteronomy, and bring into sharper focus one of the many harms of oppression.

BIBLIOGRAPHY

Baier, Annette. "Trust and Antitrust." *Ethics* 96, no.2 (1986): 231-260.

Berofsky, Bernard. *Liberation from Self.* Cambridge: Cambridge University Press, 1995.

Benson, Paul. "Feminist Intuitions and the Normative Substance of Autonomy." *Personal Autonomy* ed. James Stacey Taylor. Cambridge: Cambridge University Press, 2005.

---. "Free Agency and Self-Worth." *The Journal of Philosophy* 91, no.12 (1994): 650-668

Christman, John. "Autonomy and Personal History." *Canadian Journal of Philosophy* 21, no.1 (1991): 1-24

---. "Liberalism, Autonomy and Self-Transformation." *Social Theory and Practice* 27, no.2 (2001): 185-206

---. "Autonomy, History, and the Subject of Justice." *Social Theory and Practice* 33, no.1 (2007): 1-26

Cuypers, Stefaan. *Self-Identity and Personal Autonomy.* Aldershot: Ashgate Publishing Limited, 2001.

Dillon, Robin. "Towards a Feminist Conception of Self-Respect." *Hypatia* 7, no.1 (1992): 52-69

---. "Self-Respect: Moral, Emotional, Political." *Ethics* 107, no.2 (1997): 226 -249

Ekstrom, Laura Waddell. "A Coherence Theory of Autonomy." *Philosophy and Phenomenological Research* 53, no.3 (1993): 599-616

Frankfurt, Harry. "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68, no.1 (1971): 5-20

---. "Identification and Wholeheartedness." *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* ed. Ferdinand Schoeman. Cambridge: Cambridge University Press, 1987.

Fricker, Miranda. *Epistemic Injustice.* Oxford: Oxford University Press, 2009.

Govier, Trudy. "Self-trust, Autonomy, and Self-Esteem." *Hypatia* 8, no.1 (1994): 99-120

Jaworska, Agnieszka. "Caring, Minimal Autonomy, and the Limits of Liberalism." *Naturalized Bioethics* eds. Hilde Lindemann, Marian Verkerk & Margaret Urban Walker. Cambridge: Cambridge University Press, 2009.

Korsgaard, Christine. *Self-Constitution: Agency, Identity, and Integrity.* Oxford: Oxford University Press, 2009.

Lehrer, Keith. *Self-Trust: A Study of Reason, Knowledge, and Autonomy.* Oxford: Oxford University Press, 1997.

Luker, Kristin. *Taking Chances: Abortion and the Decision not to Contracept.* Berkley: University of California Press, 1975.

McLeod, Carolyn. *Self-Trust and Reproductive Autonomy.* Cambridge: The MIT Press, 2002.

Mele, Alfred. *Autonomous Agents: From Self-Control to Autonomy.* Oxford: Oxford University Press, 1995.

Meyers, Diana. *Self, Society, and Personal Choice.* New York: Columbia University Press, 1989.

Noggle, Robert. "Autonomy and the Paradox of Self-Creation: Infinite Regresses, Finite Selves, and the Limits of Authenticity." *Personal Autonomy* ed. James Stacey Taylor. Cambridge: Cambridge University Press, 2005.

Parfit, Derek. *Reasons and Persons.* Oxford: Oxford University Press, 1984.

Sartre, Jean-Paul. "Existentialism as a Humanism." Trans. Philip Mairet *Existentialism from Dostoyevsky to Sartre* ed. Walter Kaufman. New York: Plume, 1956.

Stoljar, Natalie. "Autonomy and the Feminist Intuition." *Relational Autonomy* eds. Catriona Mackenzie & Natalie Stoljar. Oxford: Oxford University Press, 2000.

Watson, Gary. "Free Agency." *The Journal of Philosophy* 72, no.8 (1975): 205 -220

Williams, Bernard. "Ethical Consistency." *Essays on Moral Realism* ed. Geoffrey Sayre-McCord. New York: Cornell University Press, 1988.

Wolf, Susan. "Sanity and the Metaphysics of Responsibility." *Personal Autonomy* ed. James Stacey Taylor. Cambridge: Cambridge University Press, 2005.