ADAPTIVE PREFERENCE FORMATION & AUTONOMY: MOVING TOWARDS RESPECT

Kryssa Karavolas

A thesis submitted in partial fulfilment
of the requirements for the degree of
Bachelor of Arts (Honours) in Philosophy,
The University of Sydney, May 2017

Acknowledgments

Dr. Luke Russell, for always affording my ideas respect.

My honours cohort, particularly Emily Cook, whose feedback and encouragement made this experience less daunting.

My Mother, who reminds me to never take these sorts of things too seriously.

And Adam, without whom, none of this is possible.

Contents

> "She was so evidently the victim of the civilization which had produced her, that the links of her bracelet seemed like manacles chaining her to her fate."
>
> - Edith Wharton, *The House of Mirth*

This thesis seeks to primarily answer the following question: are adapted preferences autonomous? In pursuing the answer of this question, I am unsurprisingly faced with two importantly related queries: firstly, what actually is adaptive preference formation? And secondly, what kind of theory of autonomy is correct and why?

In the spirit of question answering, the first chapter of this thesis seeks to provide a more robust account of adaptive preference formation (herein APF), a theory which states that the preferences held by an agent can be subconsciously causally produced by the restriction of options. Through an examination of Jon Elster's original account of the concept, and a consideration of Amartya Sen and Martha Nussbaum's contemporary interpretations of Elster's account, I intend to flesh out the mechanics of APF, considering the necessary and sufficient conditions for APF. This section aims to solidify the descriptive literature of APF, with a focus on differentiating the process from other similar concepts such as character planning and internalised oppression (herein IO). Ultimately, I conclude with a variation of Elster's account and produce my own examples of agents who hold adapted preferences (herein AP).

In Chapters 2 and 3 I consider the claim made by Elster and Nussbaum that AP are, by hypothesis, not autonomously held preferences. In order to adequately discern whether this statement is true, I dedicate some time to determining what theories of autonomy these philosophers subscribe to and why. I discover that they commit themselves to a substantive account of autonomy and propose arguments against their accounts. Instead, I endorse a procedural threshold account of autonomy, with the fundamental requirement that an agent reflectively endorses their preferences. With this account in mind, I move on to answer the question of whether agents can be autonomous in holding an AP. I find that – on my account of autonomy – there is no reason to believe that agents who hold AP are not autonomous. I

conclude with addressing the persisting concern that there are no longer grounds on which to scrutinise and override APF if they are to be considered autonomous. So, I make an appeal to balancing respect for autonomous preferences with the need to overcome the often oppressive conditions that cause AP.

Before moving on to Chapter 1, I will make some general comments about my methodological choices in this thesis, my reasons for pursuing this project, and situate the problem in the broader context of philosophical enquiry. Many philosophers take cases of APF to be doing political work that highlights the issues with accepting social norms simply because individuals comply with them (Terlazzo 2016). In particular, the cases of most interest to philosophers are often those that involve oppressed people endorsing oppressive norms. This is in part because they bring concerns of responsibility and respect to the forefront, but also because they highlight the persistent problem we face in understanding some agents as the *source* of their own actions, much like in cases of akrasia and addiction.

With this broader context in mind, I believe that cases of APF force one to consider how best to deal with oppression more generally. The traditional approach of dealing with APF has been to appeal to objective standards of the good in theories of autonomy to determine whether the content of an agent's preference undermines their autonomy. A lack of autonomy might then warrant intervention or withdrawal of interpersonal and political respect for the AP. Another avenue available in dealing with oppression is to take a procedural approach to autonomy, which gives us a mechanism through which to criticise the *ways* in which the preferences are formed in order to determine whether they're autonomous. This mechanism operates without reference to objective conceptions of the good life. This is the approach I plan to take in my thesis. I believe dealing with APF and autonomy in this way is strategically better for the following reasons: it avoids paternalism by ensuring it is not the case that members of oppressed groups are always deemed not autonomous; evades conflict over which objectively valuable norms to endorse; and more accurately maps our intuition about what kind of behaviour counts as autonomy precluding, such as severe manipulation or an inability to self-reflect.

This methodological consideration does not suggest that there is no set of objective universal norms that ought to be upheld (nor do I want to commit myself to this kind of view), but it does mean that seeking to justify these norms might get in the way in dealing with the unique problem of APF and theorising about autonomy. Moreover, this methodological choice does not mean that I intend to disregard the motivation for calling APF an autonomy precluding process – there must be some grounds on which to scrutinise or override preferences of these kinds; but as will be shown through my discussion of autonomy, we must seek an alternate

justification for interference, without relying on AP always being not autonomous. This alternate means of overriding might involve a reliance on objective values, but is carefully balanced with a respect for autonomous AP.

In this chapter I provide an account of APF, clarifying what constitutes an AP and how the process of adaption occurs. I begin with an analysis of Elster's original account of APF and distinguish it from other closely related concepts such as regular preference formation and character planning. I go on to analyse the contemporary interpretations of this account by Sen and Nussbaum, with a particular focus on assessing the examples they provide of agents who hold AP. I do not take either account to be completely correct with regard to what APF is, and through my analysis aim to provide the necessary and sufficient conditions of APF. I undertake this project in part because the literature on APF is often vague with regard to exactly what counts as an AP, and also because this foundational work is necessary for the ensuing discussion on autonomy and respect.

1.1. *Conceptualising adaptive preferences: Aesop's Fox Fable*

Jon Elster's account of APF begins with an analogy of a fox from an Aesop fable. The fable alone does not account for the exact process of APF, but provides a nice rubric from which to begin, and a starting point to elicit our intuitions about the kinds of cases Elster has in mind when he is describing APF.

The fox used to desire and eat grapes until they became unavailable to him – at some point, he stopped having access to the grapes because they grew on a tree branch that he could not reach. The fox finds himself no longer preferring the grapes and instead judges that they are sour and that he doesn't want them anymore. This is what Elster takes to be an instance of APF. APF happens when an agent's preferences adapt in response to circumstance. Specifically, the preferences that they hold must be causally produced by the agent's belief about the availability (or unavailability) of the potential options.[1] That is, the fox no longer preferring the grapes is the result of the grapes no longer being within reach (Elster 1983, p. 110-142).

One issue worth considering in more detail is that this fox analogy alone doesn't do much in the way of clearly differentiating AP from the normal and expected formation of preferences. All preferences are inevitably adapted to some degree because they are the result of the restricted set of options given to an agent at any given time. No agent has a set of infinite options, nor can

[1] It is worth mentioning that a restriction of options might have the opposite effect on the fox; he may come to prefer the grapes because they are no longer available. This is similar to the classic example of unrequited love, where an agent's rejection enhances their preference further. Although these are interesting questions worthy of pursuit, and prima facie possible on this account of APF, they are not included in the scope of this thesis.

they have a set of options that are identical to another agent's options. To illustrate this point, consider how I don't currently have the option to do *any possible* action. I can't get in to a spaceship and launch myself to the moon; I can't finish this paper in the next fifteen minutes; I can't do a double back flip (or a single back flip for that matter). The point I'm trying to make here is that we are inevitably constricted by the physical world and our capacities as agents. Moreover, there are no two agents who have access to exactly the same options. Even siblings with identical upbringings, access to the same resources and equal opportunity will have differing sets of options. These examples illustrate that all agents reason to their preferences from a point of restriction and these restrictions vary depending on circumstance. For the sake of distinction throughout the chapter, call these kinds of cases instances of *regular preference formation*.

Elster attempts to combat this concern by differentiating between APF and regular preference formation through introducing the idea of the "feasible set" (1983 p. 123). The feasible set is the range of options within reach for the fox; the restriction of these options is what is causally responsible for the fox's AP. In this way, AP are formed when the agent is denied some other option and genuinely believes, as a result of this change in circumstance, that they no longer prefer the unavailable option.[2] The fox holds the belief that the grapes are sour and it is this belief causes him not to prefer them. The fox doesn't contemplate or lament his restricted options, nor does he think that given greater access to the grapes that he might in fact prefer them. Ultimately, the fox's feasible set is what causes his new preference for food other than the grapes. Moreover, in APF there is no awareness by the agent that the restrictions they face might influence their preference. Elster puts this in terms of the process of adaption happening "behind the back" of the agent (1983 p. 118). So, the feasible set is also conceived of as something that the agent is not aware of.

In instances of regular preference formation, it is clear that most agents exhibit some kind of awareness that their preferences and tastes are causally produced by their lived experience, restrictions and circumstance. It is not simply the case that they come to prefer X over Y without any recognition or acknowledgment whatsoever that their personal history has played some role in determining their preference. In this way, there is no 'feasible set' in regular APF, but simply a set of restricted options that the agent is aware of, which inform their preferences. An example of regular preference formation is food preferences. For instance, I don't prefer spicy foods.

[2] It is important here to note the distinction between preference and desire. Desires are traditionally understood as addressing only one single object or goal with absolute utility whereas preferences are comparative in nature. Framing APF in terms of preferences instead of desires is not necessary but assists with the conceptualisation of APF – the agent's desires are seen as preferred over the alternative set that, given the absence of restriction, could have informed their desire.

However, my reasoning capacities as an agent are not so constrained by a lack of awareness that I can not recognise the potential of preferring something else. A way of framing this might be to consider if an agent, when probed about their preference, could imagine or consider how their preferences are responsive to their personal experience and restrictions. So, if my personal history were different and I had been brought up eating spicier foods from a young age, perhaps I would have developed a taste for spicy food. More importantly, if I were asked about my preferences, I'd be able to give a causal explanation of the reasons why I don't like spicy food, and this could include some kind of recognition that I might have developed a palette for spicy foods had I grown up with them. The important take away from this distinction is that in instances of APF the agent is never *aware* that their restricted options cause their preference and the agent's actual available set is all that is feasible for them.

Additionally, Elster spends considerable time differentiating APF from other closely related concepts where adaption appears to occur. A distinction is made between an agent's pre-commitment to a course of action and AP (1983, p. 117). A pre-commitment might involve an agent relinquishing their autonomy in the future for their benefit in the long term – a classic example of this is an agent's commitment to have faith in God and follow certain religious scriptures. The pre-commitment might prevent the exercise of autonomy in the agent's future because they have committed themselves to a certain course of action. A nun or a monk who opts in to life long religious practice with their full autonomy could be stripped of their capacity to exercise their will at a later date in a particular situation due to the requirements of the lifestyle they've chosen.[3] Another kind of pre-commitment takes the form of what Elster calls "character planning" whereby an agent has "meta-preferences" about how they wish to act, the course they intend their life to take and so on (1983, p. 118-9). An example of this is the alcoholic who gives a trusted friend the key to their liquor cabinet and insists they don't return it under any circumstances. The restriction of potential preferences is voluntary in order to achieve the agent's more important preference of sobriety over the short term preferences to ease the addiction by drinking alcohol. This kind of pre-commitment involves a conscious continual restriction of preferences in order to achieve the meta-preferences one has planned for themselves.

So, according to Elster, APF consists of the following important elements: the adaption must occur behind the back of the agent so that they are not aware that restricted options has causally produced their preference (this is primarily what distinguishes APF from regular preference formation); the preference the agent holds must be caused by the unavailability of other

[3] Whether the particular acts of the agent after their pre-commitment are autonomous despite appearing to restrict the agent's will is subject to debate but will not be discussed here.

options; instances of pre-commitment and meta-preferences, where restricted options may appear to have casually produced the preference held by an agent are not considered APF because they fail to meet the previously listed conditions.

1.2. *The contemporary account: Sen and Nussbaum*

Sen and Nussbaum offer contemporary accounts of APF that ground the common understanding of the concept in broader literature. Their discussion of APF is primarily used as an example to respond to utilitarian accounts of assessing subjective welfare. Subjective welfare is an assessment of welfare grounded in the testimony of agents about their own wellbeing. It is based solely on what the agent believes about the way their life is going. These agents might appeal to their everyday enjoyment of life, their health, happiness and aesthetic appreciation to justify their accounts of how they perceive their lives to be going. This kind of account is partly justified because it "capture[s] an important feature of the concept of intrinsic value to say that what is intrinsically valuable for a person must have a connection with what he would find in some degree compelling or attractive" (Railton 1986, p. 9). However, subjective accounts of welfare are incredibly difficult to assess for two main reasons: first, they rely on differing 'markers' of wellbeing between agents (i.e. they differ interpersonally because agents take different things to be critical to their wellbeing; for example, one agent might take health to be critical to their wellbeing while another agent takes knowledge to be essential); and second, the assessment relies solely on the agent's subjective account of their wellbeing, which is subject to adaption, manipulation and failures of rationality.

The second concern is where Sen and Nussbaum's criticism of subjective welfare emerges. They identify the difficulties in assessing agents who hold AP because the agent's capacity to analyse their own wellbeing has been compromised by APF. Sen argues that welfare is notoriously confusing terrain because an agent's expression of their "primitive feelings" (1990, p. 126) regarding their own judgment about their well-being can be inhibited by their incapacity to be self-reflective. Sen writes that "personal interest and welfare are not just matters of perception; there are objective aspects of these concepts that command attention even when corresponding self-perception does not exist." (1990, p. 126) That is to say, theories of welfare should not be grounded in an agent's evaluation of their own wellbeing alone for two reasons: firstly, agents can be wrong about their own wellbeing; and second, there are more effective objective markers for wellbeing that go beyond the subjective accounts given by agents.

Sen and Nussbaum provide a number of case studies of agents who they take to be holding AP to demonstrate how accounts of subjective welfare can be problematic for utilitarian calculus. Sen cites a survey held by the All-India Institute of Hygiene and Public Health in Calcutta after the Great Depression in 1944 (Sen 1987, p. 52-3). He found that 45.5% of widowers (males) said their health was either 'ill' or 'indifferent', whereas only 2.5% of widows (females) said they were 'ill' while none marked their health as 'indifferent'. Sen noted that this was in stark contrast to the realities of the health conditions of female widows as they were traditionally the worst off group in society and lacked basic nutrition. So, this is to say that the widows were incorrect with regard to the reality of their circumstances – they saw themselves as healthy when they were, in fact, not. Nussbaum reiterates this point with her examples of Indian women who appear to hold AP. Her first example is Vasanti, a woman who remained in an abusive relationship because she believed it was her "lot in life" (2001, p. 68) and that it was simply the reality of a woman's role to be subject to this kind of behaviour. Her second example is Jayamma, a victim of unequal family income sharing and wage discrimination who accepted that she was paid less for "heavier work in the brick kiln and denied chances of promotion" (2001, p. 69).

If agents can be wrong about their wellbeing and interpersonal comparisons of welfare are inaccurate, then there needs to be some objective marker that allows for agents' welfare to be assessed. Nussbaum provides these objective markers in her 'Capabilities Approach' and takes access to the following to be essential for human flourishing: life; bodily health; bodily integrity; senses, imagination and thought; emotions; practical reason; affiliation; other species; play; and control over one's environment (2006).[4] In light of the list of capabilities, Nussbaum takes Vasanti's and Jayamma's preference formation to hinge upon not having access to the aforementioned capabilities. So, the preferences for abuse and wage discrimination are causally produced by a lack of options and are therefore adapted.

I take issue with some of Sen and Nussbaum's cases of APF, so I will spend some time unpacking whether they are appropriately and accurately described as examples of APF, or if they are simply useful examples to prove that subjective welfare is flawed. First, a necessary element of APF is that the agent *prefers* the treatment that they face. What does it mean to 'prefer' in the context of APF? Preference doesn't necessarily entail that the agent is enthused by their situation, nor need it mean that agents actively seek out the preference with which they end up. Instead, it minimally requires a lack of protest to the situation agents find themselves in. In light of this understanding of preference, I think that an intuitive response to Sen's Calcutta widows example and Nussbaum's Vasanti is that they don't fit this characterisation of APF, because it appears that

[4] Whether these capabilities are objective goods is far from settled and I will not take issue with this problem here.

the agents don't 'prefer' anything, they are simply wrong about their conditions. In Sen's example, the widows are incorrect about the state of their wellbeing and make an unfounded judgment about their welfare. The widows do not prefer their illnesses or poor nutrition over health; they are just unaware of the gravity of the suffering they face or that they are facing it all. It is difficult to see how Sen's example amounts to an AP when the requisite element of the agent having a 'preference' is missing. Alternatively, perhaps if the widows had made this judgment and actively expressed their preference to remain malnourished, then they might hold an AP.

Nussbaum's example of Vasanti may not appear to be an instance of APF because Vasanti's behaviour in response to the abuse she faces could be interpreted as amounting to an objection or protest to her situation. Vasanti's responses when probed about her situation is that it is her "lot in life" (2001, p. 68), which seems to suggest that she doesn't actually enjoy being with her husband but believes she is stuck there for reasons out of her control. Although this critique of Nussbaum's response seems warranted, it is a very uncharitable reading of Vasanti's case. Although Vasanti mightn't prefer her situation in the strong sense (insofar as she doesn't express joy or excitement), she prefers it over the alternative of divorcing her husband. In turn, this belief results in the prospect of staying in the marriage the object of her preference. We can see how this quite closely follows Elster's fox fable rubric of APF: an option is denied to the agent (either the grape is too high up or divorcing your husband isn't an option), so in turn a preference is developed for the alternative (that is, the food within reach or to remain in the abusive relationship).

Nussbaum's overarching commentary about her examples of APF is that the women "lacked the sense that what was happening was wrong" (2001, p. 69) and failed to appreciate their entrenched rights to be free of abuse, have equal wages for equal work and so on. Due to the restricted options agents face, they are unaware that their contexts are detrimental to their wellbeing. However, as we've seen in the case of the Calcutta widows, the condition of a lack of awareness of wrongdoing done against you isn't sufficient on its own for the example to count as APF. So, it isn't enough that the agent "lacked the sense that was happening was wrong" (2001, p. 69), they must also hold a preference for something over an alternative.

1.3. *Differentiating between AP and internalised oppression*

The discussion so far might have stimulated the concern that APF is simply an example of internalised oppression (herein IO). In this section I argue that they are distinct concepts that often overlap and that some clarificatory work should be done to pull the two apart as it assists in clarifying APF conceptually.

In feminist theories of autonomy, IO is conceptualised in a multitude of ways. The classic case takes the following form: an agent who is a member of an oppressed group internalises the oppressive ideology they are subject to by taking on board the false stereotypes associated with the group, and goes on to develop desires, preferences and behaviours in line with the expectations of the stereotype. This is done either by actively championing the kinds of behaviours expected of them or by simple acceptance of the conditions without protest. However, as I will prove in this section, simply because the agent has a general pro-attitude towards oppressive conditions, this does not mean they have the requisite causal story for the preference for it to count as APF. Some instances of IO are not behind the back of the agent, nor is the restriction of options that occurs in oppressive circumstances always the actual cause of the preference that the agent holds. Additionally, APF can occur to people who are not members of an oppressed group, so not all cases of APF are going to be cases of IO.

Before considering the aforementioned differences, I will sketch a case of IO to make the distinction easier to identify.

> SARAH has been raised in a household that values and promotes female domesticity. She admires her mother, who embodies these values by staying at home and performing domestic duties. It is also the case that Sarah has been subject to traditional narratives about women being domestic and not pursuing education or a career. Sarah prefers to stay at home and raise a family, and does not intend to pursue tertiary education.

Initially this example might appear to be a case APF because the process of internalisation could amount to an actual restriction of options that causally produces a preference of not wanting to pursue an education or career. It could be that the stereotypes Sarah is exposed to restricts her potential options. That is, through oppressive narratives that she subconsciously internalises, the oppression acts to restrict the options Sarah considers herself to have. However, if this is the case, then this example should show that not only does Sarah not want to be educated, but there was no potential option for her to be. However, the jump made from oppressive stereotypes to restricted options here is tenuous and not particular convincing. Although dominant social narratives inevitably affect the behaviours of members of that social group, the narratives themselves do not necessarily close off options in the same way literal restriction of options do.[5] Additionally, the reasons for Sarah preferring not to pursue tertiary education is not clearly causally related to the supposed restrictions placed upon her. There are more obvious alternate

[5] Oppressive stereotypes are often overcome by marginalised individuals through the express preference for something that isn't available to them. E.g. female suffragettes who protested for their enfranchisement.

explanations for her choice, even if she doesn't see them as reasons. For instance, admiring her mother might be an alternate plausible causal explanation.

This does not mean that agents subject to IO are never victims of APF. In fact, I believe it is fair to maintain that many instances of IO are also examples of APF. The example of Vasanti from Nussbaum amounts to both IO and APF. The process of internalising stereotypes about women being at the behest of their husband can also be characterised as a response to the restricted options available to women in her context. This suggests that perhaps there was no viable option[6] for Vasanti to be anything other than in an abusive relationship. In her case, the stereotypes that are reinforced by the dominant social narratives about women's roles restrict her options. Moreover, her preference is directly causally produced by these restricted options, which are the result of the oppressive conditions, and there are no plausible, convincing alternate causal explanations as to why she holds her preference.

To conclude, not all agents who hold AP are members of an oppressed group (e.g. the fox in Elster's fox fable is not part of an oppressed group), nor do victims of IO always hold AP, as we saw in Sarah's case. However, the kinds of APF that are of interest to philosophers are often cases of APF that are formed under oppressive circumstances because they reveal an oddly formed preference that is supposedly problematic due to its contents.

14. *Limits of the fox fable*

In the last few sections I have demonstrated that the contemporary accounts of AP from Sen and Nussbaum have introduced cases that are not discussed by Elster in his initial conception of the fox fable. In this section I will explore these points of difference and defend a broader account of APF that doesn't adhere strictly to the fox fable rubric.

In the case of the fox, the APF happens after a *change* in the available options has occurred. That is, the fox once had access to the grapes and no longer does; the available options have decreased. On the other hand, in the cases of Vasanti and Jayamma, the agents' options have been restricted from birth. The adaptive preference does not occur after a change in circumstance. Instead, the agents have never known any different and their adaptive preference is a result of life long habituation. Elster makes a point of stating that AP occur after a change in the options available to the fox (1983, p. 133) but provides no explanation of why there must be a change. Perhaps Elster makes this point salient because he believed the change occurring during the agent's

[6] For my purposes in this thesis, I consider a viable option to be either an actual option that is available to the agent, which they are aware of, or an option that has been entertained as possible in the imaginary of the agent (insofar as it has entered their conscious thoughts as a possible option).

life highlights the peculiarity of APF. However, I do not think Elster has provided adequate reason to restrict a theory of APF to during-life changes in options. The shift from preferring X to not preferring X is not clearly important to the integrity of the concept on Elster's account, nor does it change the actual process of adaption. What seems to be of concern is that the restricted options cause the preference, not that the options are restricted at some given time.

Nussbaum has argued that another way Elster's conception of APF is limited is that it doesn't allow for the possibility of positive, prudentially good adaption (2001, p. 78). As Nussbaum has noted, there are countless cases where being sensitive to restrictions allows agents to set realistic goals for themselves and adjust their preferences so that they can be satisfied in the future (2001, p. 78-9). In this way, APF yields positive outcomes and allows agents to be more realistic about the course their life will take. Nussbaum provides an example of a child who, at a young age, dreams of flying (2001, p. 78). As they grow older and become aware of the restrictions they face, they come to realise that this is not a possible option – that is, their preference to fly is changed as a result of the option of flying not being available to them.

A query worth raising in response to Nussbaum's point is whether this adaption occurs behind the back of the agent in the way that Elster requires. I think that this is hard to say with certainty – there are multiple scenarios where the child may or may not be aware that APF has occurred. For instance, perhaps later on in life the child critically reflects and recognises the shift that has been undertaken in response to restricted options, but during the period where her preferences changed, she was unaware and simply undergoes a process of adaption in response to the options given to her. Alternatively, maybe the child is told about their restricted options, so they adapt their preferences in light of new information and a set of newly restricted options. No matter what story you tell about the child who wants to fly, it is plausible that at least one account allows for the conclusion that she no longer wants to fly because of restricted options that she is not consciously aware of. If we accept this, then we can be sympathetic towards Nussbaum's concern that Elster was wrong to characterise all instances of APF as negative.

1.5 *Necessary and sufficient conditions for APF and revised examples*

Now that I have spent considerable time analysing the literature on APF and refining the concept, I will set out what I believe to be the necessary and sufficient conditions for APF. I will then go on to provide two revised examples of agents who hold adaptive preferences. I do this because the concept of APF is seriously under-formalised in the literature and knowing exactly

what counts as and constitutes an AP allows for a more accurate discussion of whether APF is autonomous in Chapters 2 and 3.

Necessary and sufficient conditions:

1. The agent is subject to a set of restricted options. These options needn't have been restricted or altered in the course of the agent's life; the options can be restricted before birth. This restriction can be literal or imaginative.
2. The restricted options *cause* the agent to form a particular preference for an option that is available to them. That is, the preference that the agent ends up with is causally produced by the fact that they are restricted in their options.
3. The adaptive process (1 & 2) happens behind the back of the agent. Viz., the agent is not aware that their options are restricted, nor are they aware that their preference is a result of the restricted options.

Other unnecessary but common elements of APF:

4. In instances where the agent is harmed or detrimentally effected by their AP, they often do not see themselves as the victim of oppressive or adverse circumstances. This is a common but not universal outcome of the third necessary condition listed above.
5. APF and instances of IO often overlap as victims of IO tend to develop AP that fulfil the required conditions set out above. This is not always the case, as we saw in 1.5.

Now that I have set out the basic conditions for APF, I will spend some time formulating what I believe to be two cases that exemplify the above necessary and sufficient conditions, inspired by Nussbaum's real life cases of agents she takes to hold AP (2006).

JAYAMMA is a victim of unequal family income sharing. She is paid less for doing the same amount and same quality of work as her male co-workers. Moreover, when she takes this income home to her family, she is expected to give it all to her husband. She has been subject to these expectations since birth. She has a restricted set of options insofar as the possibility to be paid the same as her male counterparts is not a tangible possibility nor entertained by Jayamma herself; the prospect of not giving up her income to her husband is not an option given the expectations of women in her society and what she individually perceives to be viable. Additionally, when probed about the situation, she states that she prefers this arrangement compared to one in which she keeps her money and earns the same amount as her male colleagues, and sees no problem whatsoever with the conditions she is subject to.

JO prefers to stay in an abusive relationship because divorcing her partner is not a viable option. The set of options she has is restricted as a result of her own belief in the roles of women in marriage, which was informed by social discourse and her upbringing. Jo is not aware that she holds these views about marriage and that they inform her preferences in relationships. Her beliefs prevent the available set of options to include separation or divorce. Jo accepts the abuse as permissible, seeing no problem with her treatment. Moreover, she has a preference to stay in this relationship with her spouse and has no intention of leaving them.

It is important to note that in these cases the preferences held are caused by the restricted options they face and not some alternate causal explanation as in instances of IO, and that this process of adaption happens behind the agents' backs. These examples align with the conception of AP I have provided in this chapter, encapsulating what I believe to be a more thorough theoretical framework of APF.

This chapter is dedicated to discerning what accounts of autonomy both Elster and Nussbaum endorse when they make the claim that APF is not autonomous. I discover that they endorse a substantive account of autonomy, and provide arguments against theories of this kind.

2.1. *Elster's theory of autonomy*

Elster's account of APF presupposes that agents who hold AP are never autonomous with regard to their preference. Elster's conception of APF states that the agent must not be autonomous with regard to their preference formation in order for the preference to count as adapted. This is to say that Elster's account of APF has a *necessary condition* that the agent is *not autonomous* in their preference formation if it is to count as an AP (1983, p. 132-3). Moreover, Elster states that his theory of "autonomy will have to be understood as a mere residual, as what is left after we have eliminated the desires that have been shaped by [APF]." (1983, p. 24).

Elster's conception of autonomy appears to have two main elements: first, a requirement that the agent has "the freedom to do otherwise" (1983, p. 129) (I will call this the *freedom* condition); and second, that the agent has a requisite level of rationality with regard to their desire (I will call this the *rationality* condition). These are very vague conditions, so before discussing their validity in constituting autonomy I will undertake an interpretive project to figure out what Elster means.

First, he confesses that he will not give a "positive characterization" of autonomy (1983, p. 129). However, I think we can get closer to ascertaining a positive account through exploring his explanations of the aforementioned conditions of *freedom* and *rationality*. He states that autonomy is a "substantive rationality of desires … being for desire what judgment is for belief" (1983, p. 30). Here we start to see an argument by analogy emerge; Elster writes that conceptualising autonomy "rests on [an] analogy with judgment" (1983, p. 129). However, we are restricted in interpreting the analogy, even if it is valid, because Elster hasn't provided an explanation of "substantive rationality". Elster's main means of getting us closer to defining substantive rationality is by differentiating it from regular 'thin accounts' of rationality (1983, p. 15). Thin rationality merely requires internal coherence and consistency in the mind of the agent. He describes substantive rationality as something that goes "beyond the exclusively formal considerations" (1983, p. 5) of coherence. This suggests a reliance on some external standard of

rationality, perhaps grounded in normative evaluations made by the agent, in order for desires to be considered autonomous.

Exploring the analogy to judgment further might give us some additional clarity on the definition of substantive rationality. The analogy is intended to work in the following way: beliefs are epistemically justified insofar as they are grounded "in the available evidence" (1983, p. 17), and desires are autonomous insofar as they are based on substantive rational beliefs. Therefore, autonomy is the analogous element for desire as judgment is to belief. This is to say, no desire is autonomous without the agent's exercise of substantive rationality, just as no justified belief can exist without a judgment based on evidence. This however, doesn't seem to do much in the way of clearing up what substantive rationality means. At the very least, we can say that Elster wants substantive rationality to mean something more than internal coherence.

I turn to Colburn's interpretation for some analysis on the meaning of substantive rationality. He provides a very charitable reading of Elster and restates Elster's conception of autonomy positively in the following way: autonomy is "deciding for oneself what is valuable, and living one's life in accordance with that decision" (Colburn 2011, p. 61). Here we start to see the substantive element of rationality emerge in the form of a value judgment. For the agent, this might involve judging what desires are worthy of pursuit, what desires are valuable, and what desires are good to have. As I discussed in Chapter 1, Elster wants to steer clear of agent's subjective accounts of their own welfare, so the value judgments involved in substantive rationality can not simply be about what the agent perceives to be good for themselves (as this is one of the problems that stems from APF), but instead must also be the kinds of things that are in fact objectively good. For instance, subservience or oppression are objectively bad things, so any pursuit of them by an agent would render the agent irrational and therefore not autonomous. That is to say, the object an agent pursues must be substantively good for the preference to be autonomous. This commits Elster to a very strong substantive account of autonomy that I will criticise in the next section.

Colburn's re-characterisation of Elster's substantive rationality principle also inadvertently deals with what I identified as the second condition required of autonomy on Elster's account: *freedom*. The "decides for oneself" element speaks to Elster's condition of the agent's "freedom to do otherwise" that he believes is absent APF, which is a paradigmatic case of non-autonomous desires. In deciding for oneself on Elster's account, the agent must "want to do x, […] is free to do x, and free not to do x" (1983, p. 130-2). So, for an agent to be truly free with respect to their options, they must have the freedom to choose between options. This means that, in order for a desire to be autonomous, it cannot be shaped by necessity and must be decided upon by the agent

in an unrestricted way. Elster's account of freedom with regards to autonomy seems to rely heavily on the choice to prefer something, suggesting that freedom is linked to the options that an agent has. If an agent has access only to restricted options, as is the case with AP, then they are not free to prefer not-X (and not free in the way Elster requires) and therefore they are not autonomous.[7] So, Jayamma is not free to prefer unequal wages because she is not free *not* to prefer it.

Elster's account of autonomy, although at times confusing and unclear, can be formulated positively, without reference to APF, in the following way:

> Autonomy is exercised when an agent *rationally* decides what they desire, *free* from restriction and coercion, and live their life in *accordance* with that substantively rational decision. Decisions are irrational when the content of the desire held by the agent is objectively not valuable for the agent; often these substantively 'bad' desires are developed as a result of a lack of freedom.

Immediately we can see a lot that is wrong with this account of autonomy; in particular, a persisting vagueness that makes autonomy difficult to assign to any agent and a concern that ideas of substantive good are difficult to articulate. However, before considering why Elster's autonomy is incorrect, I will consider whether Nussbaum's account of autonomy aligns with my positive formulation of Elster's theory.

2.2. *Nussbaum's theory of autonomy*

Nussbaum appears to implicitly accept Elster's account of autonomy by accepting his definition of APF with little criticism. Nussbaum's only change to Elster's APF, as we saw in Chapter 1, was her addition of life long AP and the potential for positive adaption. These critiques of his account were justified, as discussed earlier. The concerns expressed by Nussbaum hardly amount to strong criticisms against Elster's account as they do not pose a direct challenge to his rubric of APF, nor does it undermine what I concluded to be the necessary and sufficient conditions of APF in 1.5. This suggests that Nussbaum endorses the accompanying "residual" account of autonomy from Elster (1983, p. 24)

The fundamental sentiments of Elster's theory of autonomy are echoed throughout Nussbaum's work. For instance, Nussbaum suggests that the difference between an agent who holds AP and one who doesn't is the "difference between fasting and starving" (1999, p. 44). Much

[7] Note that there is a difference between the freedom to *prefer* X and the freedom to *do* X (in other words, the freedom to act upon a desire). Presumably there are cases where an agent can prefer to do something but is not free to act upon it. For instance, an agent might prefer to leave a marriage but is incapable of doing so. The discussion of autonomy in my thesis relates to the *preference* to leave the marriage, not the *ability* to act upon this desire.

like our previous discussion of freedom and substantive rationality on Elster's account, it appears as though the fundamental difference Nussbaum points to between agents who are autonomous and those who aren't is the kind of difference exemplified between an agent who fasts and an agent who starves. The difference between them is the freedom to choose not to eat. Freedom, as we discussed in 2.1., involves a freedom to choose what you *want* and a freedom to choose *otherwise*. To illustrate this point further, the dichotomy between the starved and the fasting highlights that the starved person had no choice to eat food and therefore was not free with respect to their choice about starving, even if they purport to be free and desirous of their own starvation, as would be the case in APF. Therefore, at the very least, Nussbaum wants to elicit intuitions similar to what Elster's account does in suggesting that in order for preference to be autonomous, it must be free from restriction or coercion, and there must be the freedom to not prefer it.

Nussbaum also argues that the way to fix problems such as APF, and perhaps something that operates independently as a marker of autonomy, is agents' access to the range of 'capabilities' she outlines in her Capabilities Approach. She believes that her list of capabilities should be accessible to all agents. On her account, access to the capabilities will increase agents' range of valuable options, which broadens the restricted feasible set that causes AP, thus giving rise to autonomy. So, she maintains that the fasting agent is fundamentally different than the starving woman because the former has choice in the matter and the latter has no "real opportunity" to not starve (2010, p. 10). This incapacity violates the capability of bodily integrity on Nussbaum's capabilities and therefore renders the agent non-autonomous. So, not only does Nussbaum require freedom of agents in order for them to be considered autonomous, she also requires that this freedom is with respect to the capabilities she deems essential. This freedom with respect to capabilities is similar to the story of substantive rationality from Elster that I told earlier, insofar as the agent must be aware of what is valuable in order for their preference be rational, and this value is justified by appealing to an objective good.

In conclusion, there are strong parallels between Elster's and Nussbaum's implied accounts of autonomy. They both require freedom with regard to valuable options. Nussbaum provides more in the way of characterising these options through her Capabilities Approach, while Elster leaves the question of what substantive rationality is open for interpretation. Both accounts seem to have two distinct more general elements worth noticing: a procedural concern that the agent is free to choose otherwise with respect to their desires, and an underlying substantive question of whether the options they have access to freely choose from are objectively good or valuable. I dedicate the next section to assessing the validity and effectiveness of substantive accounts of autonomy more closely.

2.3. *Against substantive theories*

In order to conclude whether APF is, by hypothesis, not autonomous, there must be a correct and clear account of autonomy by which to assess cases of APF. I do not accept Elster and Nussbaum's theories of autonomy. I will spend this section demonstrating that Elster and Nussbaum propose a substantive account of autonomy, then go on to argue that any substantive account of autonomy should be rejected.

Before addressing these concerns, I'd like to consider broadly what autonomy *is* and what it ought to do as an operational term we assign to agents. In moral psychology and political philosophy, autonomy is conceived of as a kind of self governance or direction that agents have with respect to their desires, actions and beliefs. To be autonomous is to act on reasons that are ones own insofar as the agent is free of coercion, manipulation or deviant causation (Dworkin 1988, p. 121-29; Arneson 1991). Metaphorically, autonomy might be conceived of being the author of one's life. A theory of autonomy should capture this sentiment while also aligning with what we intuitively believe to be paradigmatic cases of autonomous behaviour. Moreover, autonomy has value associated with it (Wolff 1970). It is a concept that, when applied to or associated with agents, attracts a kind of judgment about how that agent should be treated. For instance, a claim that an agent is not autonomous might have bearing on whether or not an agent is deemed morally responsible for their actions, whether they should be held accountable and whether other agents should respect their actions (Wolff 1970; Ripstein 1999). This is a motivating consideration when developing a definition of autonomy as, ideally, we want a theory of autonomy to track our intuitions about who and what is worthy of respect and responsibility.

I support the early feminist autonomy theorists' method of rejecting traditional notions of autonomy (Mackenzie & Stoljar 2000). These notions promoted exclusively (traditionally and stereotypically) masculine accounts of personhood (Kant 1785; Rawls 1971), including ideas that autonomous agents should be completely self sufficient, un-reliant on social relations for flourishing and have rational capacities free from distortion by emotion (Dworkin 1988; Berofsky 1995). This conception of autonomy was fundamentally problematic because it excluded traditionally feminine individuals from ever being considered autonomous. For instance, women who are financially reliant on their male counterparts for survival are not considered self sufficient, and therefore not autonomous on these accounts. Moreover, women have traditionally been perceived to be irrational or controlled by their emotions, so they are automatically deemed not

autonomous on the traditionally masculine account.[8] For a theory of autonomy to – by hypothesis – exclude such a large group in society is problematic for two reasons: first, it doesn't match our intuitions that many women are acting in accordance with their wishes and are self-governing even if they do rely on others for their livelihood; and second, it has serious implications of making it justifiable to exclude women from important social and political spheres by rendering their desires and actions not worthy of respect. These problems warrant a reconsideration of the definition of autonomy.

Emerging from this rejection is a relational account of autonomy, which attempts to re-characterize autonomy in opposition to the masculine centred accounts (Mackenzie & Stoljar 2000). Broadly speaking, there is a split in relational accounts between substantive and procedural concepts of autonomy. Substantive accounts of autonomy are value laden insofar as they require agents to achieve a particular outcome in order for their actions to be considered autonomous, whether this be to achieve wellbeing, as described in some specific theory, or always prefer things that are objectively good. Procedural accounts are content neutral and seek to theorise about autonomy without reference to substantive notions of the good. Philosophers such as Benson takes a 'strong' substantive approach to autonomy by measuring the content an agent prefers or desires against social and moral norms in order to determine whether the agent is autonomous (Benson 2005, p. 133). For instance, an agent who prefers to be free from oppression is prima facie substantively autonomous because their preference is normatively justified as oppression is objectively bad. Any agent who preferred their own oppression would be non-autonomous because a preference for an objectively bad norm precludes the desire from being autonomous. A less severe account of substantive autonomy – a 'weakly' substantive account – is another way of conceptualising autonomy (Richardson 2001). This kind of account does not constrain the agent so directly, but instead operates more abstractly through general requirements of self-respect or self-worth (Govier 1993) that have substantive elements. So in instances where agents prefer subservience, provided they are considered to have self-respect, their preference is autonomous.

The Elster-Nussbaum conception of autonomy is an example of a substantive theory of autonomy. Their appeal to a fundamentally normative constraint, either in the form of substantive rationality or capabilities, underpins their account of autonomy. The appeal to conceptions of the good can be seen in Elster's substantive rationality requirement by deeming an agent rational only when they pursue objectively valuable outcomes. Similarly, on Nussbaum's account, the

[8] This stereotype is unfounded and no less true for women than it is for men, but still pervades the social imaginary insofar as it dictates what we think about women as rational agents. This incorrect assumption fundamentally impacts whether or not women are seen as autonomous, especially when theories of autonomy are set up in opposition to stereotypical female traits.

capabilities she believes agents should have access to in order to be considered autonomous are in and of themselves constitutive of 'the good life' (Nussbaum 2000, p. 74). Failure to have unrestricted access to the capabilities means agents can never have an objectively good life, and without this, they can never be considered autonomous with respect to their preferences.

I will now spend some time proving that this kind of approach to theorising about autonomy is problematic by proposing three arguments: (1) an appeal to the good life is notoriously challenging to justify successfully (and without criticism) and therefore difficult to use in theories of autonomy. This suggests that an alternative approach to autonomy, despite whether there exists a substantive conception of the good that is in fact correct, is more useful in theorising about agents; (2) these substantive theories of autonomy require too much of the agent to be considered minimally autonomous and is contrary to common intuitions about what is considered autonomous; and (3) theorising about autonomy this way justifies paternalism in broader political and social contexts, which is something we should be seeking to avoid in philosophical discourse.

First, I will discuss how appealing to objective standards of the good is difficult and undermines the project of autonomy. This does not necessarily mean that there is no substantive conception of the good life that is objectively right, but it does provide reasons to seek an alternate explanation of autonomy. After all, autonomy is not merely descriptive – it has functions of assigning responsibility and respect, and controversy over what counts as objectively 'good' can undermine the project of making autonomy useful. Broadly speaking, conceptions of the good life are met with criticism from two different camps of philosophers: those with a competing conception of the good life; and those that think there is no such thing as the objective good, but instead endorse a kind of moral or cultural relativism (Harman 1996; Prinz 2007). Criticisms of the first kind might be seen in the disagreement about whether Nussbaum's capabilities theory or an alternate perfectionist[9] account of the good is correct with regard to what constitutes the good life. That is, there are objective values but they are difficult to pin down and to endorse across a range of individuals and societies. This kind of disagreement presents problems for endorsing and enforcing a theory of autonomy with any weight in political and social spheres. If acceptance of Nussbaum's account means endorsement of her perfectionist account of the Capabilities Approach, and there are agents who have competing conceptions of objective values, then the force of autonomy is undermined. If agents are in disagreement about what it means to be autonomous, then they would disagree about what kinds of behaviours or agents are worthy of respect, blame and so on. This same concern extends to those that hold there are no true objective

[9] Perfectionism is the philosophical commitment that underpins strong substantive accounts of autonomy; perfectionism holds there are values that are valid despite whether agents or populations endorse them (Hurka 1993, Sher 1997).

facts about values. Some philosophers hold that normative systems are constrained to their context and culture, and that an appeal to meta-justifications for our normative systems is fruitless (Sankey 2010, 2011). Whether these positions regarding objective good are correct is up for debate, but highlighting these competing conceptions of the good demonstrates how a substantive theory of autonomy is less effective in impacting the agents whose autonomy is in question. For instance, competing claims about the relevance of self-respect to autonomy makes it difficult to justify overriding preferences that exhibit a lack of self respect. Since there is no consensus on what is substantively necessary for autonomy, the status of the agent's autonomy is debatable and the treatment of the agent is contentious. From this we start to see that there are prudential reasons to accept a non-substantive account of autonomy – we ought not to rely on an account of autonomy that practically fails to fulfil what is required of autonomy as a tool. Ultimately, a theory of autonomy with no force should be avoided.

Another reason to reject substantive accounts is that they often require too much of the agent. Due to the high threshold some substantive accounts require, almost all agents fall short of being considered autonomous, and many agents that appear to be intuitively autonomous are deemed to not be. For instance, requirements of self-trust supported by theorists such as Govier (1998) require that agents must have certain self regarding attitudes and exhibit behaviours that exemplify those attitudes in order to be considered autonomous. Govier demands that one have positive attitudes towards their competencies and that they "have a general disposition to see oneself in a positive light" (1998, p. 106). For instance, a man that has no clear failings of rationality or coherence, and who is free from manipulation or coercion, could be suffering from depression that prevents his capacity to see himself in a positive light. He is excluded on the self-trust account because he doesn't see himself as capable of certain actions even when he is, nor does he trust himself to perform to a certain standard even when he is in fact capable of it. For example, the psychological work conducted on depressive realism has shown that depressed people are more likely to rate themselves as below average drivers, whereas people who do not suffer depression consistently rate themselves as skilled drivers (Burton, 2012). This seems to demonstrate that those who suffer from depression lack self-trust with regards to driving. However, this kind of failing to be self-trusting does not seem to give reason to believe the agent wasn't autonomous. It does not capture our intuitions about self-governance or being the source of ones' actions – it is not clear how the absence of self-trust precludes one from being self-governing. The agents who exhibit a lack of self-trust do not necessarily fail to meet intuitions about autonomy simply by virtue of these character traits. These 'negative' self-regarding attitudes are present in many diverse agents who we consider to be autonomous.

Finally, objective conceptions of the good usually deem agents who are already marginalised in social and political spheres non-autonomous. The issues we saw with traditionally masculine accounts of autonomy excluding socialised women is an issue that persists in substantive accounts of autonomy for other minority groups. On Nussbaum's account, any agent that does not have control over their environment is not considered fully autonomous. There are plenty of people subject to poverty who are incapable of controlling their environment due to lack of power and resources with regard to change. For instance, some agents living in Syria have no control over the war torn environment in which they currently live. If we start to single out certain groups as always not autonomous due to their lack of access to certain substantive goods, then paternalism with respect to their decisions and actions is more often justified on the grounds that they lack autonomy. This paternalism could be in the form of interpersonal social assumptions that all agents emerging from these contexts are incapable of autonomous action and thought, or it could be political in nature and take the form of legislation that directly effects the rights and responsibilities of these supposedly not autonomous agents. This not only manifests an inappropriate arrogance towards people who are members of oppressed groups, but also unjustly stereotypes agents with no justificatory evidence. These are good reasons to not accept a substantive account of autonomy.

To conclude my criticism of substantive accounts of autonomy, it seems that any theory of autonomy should aim to be descriptively accurate and match our intuitions about what autonomy, in fact, is. Substantive theories of autonomy are not only prudentially bad insofar as they are not helpful when it comes to real world arguments about autonomy and its application, but also seem to undermine our general intuitions about who and what counts as an autonomous agent. So, an appeal to an alternate account is necessary.

The alternative to substantive accounts of autonomy is a procedural account, which can be understood as a theory that commits to *not* importing substantive conceptions of the good into the conditions for autonomy. I look to the procedural accounts from Christman (1990; 2000; 2009), Friedman (1997; 2003) and Meyers (1987; 2000) to construct my own hybrid account. In particular, I intend to support a historical account of procedural autonomy that has the primary condition of self-reflection, and considers people as having varying levels of autonomy. I will also make the distinction between local and programmatic autonomy as a means of appealing to the intuition that agents who hold AP are not autonomous in the same way as agents who hold regular preferences.

3.1. *Procedural account*

Christman provides the basis for my historical procedural account of autonomy. Historical accounts can be understood in opposition to using a single, current mental state to discern whether an agent is autonomous. The intuition from Christman is roughly this: assessing agents as autonomous because they "thoroughly embrace" their current situation (values, desires, behaviours) may leave the agent open to the forces of past manipulation that we'd traditionally see as autonomy defying (2009, p. 140-5). That is, an agent can fully endorse their current state even if they were subject to a deviant causal process in the past. A failure to consider how events other than the agent's current state might lead to calling agents autonomous when they clearly aren't. For instance, I might have been hypnotised in to desiring only spicy foods at T1. A few days later, after successful hypnosis, at T2, I fully endorse my preference for chili. There's something odd about calling my preference at T2 autonomous because I have been the subject of manipulation that has, in fact, caused my preference. The historical approach from Christman attempts to rectify this problem by requiring the agent undergo a test of authenticity (2009, p. 144). This test of authenticity takes on the form of "nonalienation" (2009, p. 155). An agent's desire is authentic *iff* they were to reflect upon historical processes that led to their desire and they do not feel alienated from it. A feeling of alienation on this account is considered to be a negative judgment about or a negative emotional reaction to the desire or its formation. Nonalienation is the absence of these responses and can be thought of as a kind of endorsement. Christman notes that there are both cognitive and emotional elements that amount to a "broad consideration of aspects of oneself in light of one's history" (2009, p. 154).

One preliminary concern for this approach is that the threshold for the agent to reflect requires a kind of self-awareness that is far too difficult for most agents to attain. Agents don't usually actively or consciously self-reflectively endorse their preferences on a regular basis, nor is it something they do with regard to all desires and preferences. Perhaps it is the case that the nature of some preferences, such as the preference to marry or have children, might lead to reflecting reasons for developing them or the impact that they will have on our lives more broadly. However, most of the time we tend to act in accordance with our preferences without much consideration. In turn, we face the problem of most agents being considered non-autonomous even when this is not our intuitions about them at all. Christman responds to these concerns by highlighting that agents only have to be "minimally reflective." (2009, p. 154). Instead of requiring complete "self-transparency" (p. 154), self reflection is a kind of minimal self awareness that requires the agent to reflect upon the available evidence and causal story for accepting their values and desires. This needn't mean the agent makes special insight in to their reasons for a preference, but that they, *upon* reflection, don't conjure up negative attitudes or emotions towards the reasons for their decisions.

A second concern that emerges from this is what self-reflection should involve in order to produce the kind of nonalienation Christman is suggesting. Is it about whether the agent can reflect about the desire being what is best or what is good for them, or is it whether the agent sees the desire as authentically their own and is correct in this assessment? Christman's response to this question is that the agent need not make a correct value judgment about what is objectively best for them (if that judgment exists at all), but instead that they are simply not alienated from the desire. The latter suggestion about there being a correct authentic self discovered in self-reflection is dismissed by Christman through a general rejection that there exists an authentic self at all (2009, p. 150). He doesn't endorse a picture of the self that has, at its core, a *true* character with a set of desires and values that should exist despite the causal forces that have lead to the agent's current state. For Christman, all that it means to be authentic is that the agent is not alienated from their preferences. For it to be inauthentic is for the opposite to hold true, viz., that the agent feel alienated from their preferences. So, in rejecting the idea that there exists a single authentic desire, he also rejects the presupposition that in self-reflecting one must be correct about whether or not they feel alienated – authenticity does not depend upon the correctness of their judgment of alienation, because there is no true authentic self from which to make this judgment. So, Christman rejects both these interpretations of self-reflection and simply characterises it as a reflection that is conducted in light of an "ongoing autobiography" (2009, p. 154) – that is, reflection has to do with thinking about the desires in terms of the past that may have caused them, and the future that

contains the persisting desire. The agent is autonomous in light of this self-reflection provided it does not elicit feelings of resistance for the agent towards the desires on which they are reflecting. As long as the agent does not feel resistant to it forming a part of their future life and they are not uneasy about the history that may have caused it, they are autonomous with respect to this desire. To illustrate this point, I will provide a brief example. Imagine an agent, GEORGIA, is reflecting on her desire to engage in certain sexual behaviours. Provided she does not feel alienated from the personal history that has led up to this point of desire (that is, she feels positively about the causal processes that she deems to have influenced the development of this desire), and that she is not resistant to the desire's integration in to her future life, then she is autonomous with respect to that desire.

The next question for Christman is what do these feelings or judgments of resistance consist of? I think the answer to this is embedded deeply in the sentiment of alienation. To be alienated is to feel distant and causally distinct from something. The agent, when alienated, does not have control over the object of desire or value that they hold, but instead feels separate from it, both when looking backwards at its causes and forward as part of their ongoing life. There needn't be some finite list of feelings or judgments about the desire that amount to a resistance, but at the very least it seems that negative affective attitudes and a lack of clarity with respect to the reasons for holding the desire form the basis of alienation. Thinking back to the example of GEORGIA to demonstrate, feelings of alienation might include a confusion about why she wants to engage in these acts, or perhaps involve a negative emotional response such as sadness or anger when she considers the sexual acts taking place in her future.

I accept Christman's account of autonomy so far and believe that he has provided a solution to the issue that non-historical procedural accounts of autonomy pose, viz., that they expose agents to manipulation or deviant causation. However, I think there is a persisting intuition that some agents who are nonalienated from their desires are simply fundamentally different than their self-reflecting autonomous peers because of the kinds of things they endorse, and hence are not really autonomous. To resolve this issue, I will turn to Freidman's use of a procedural threshold account of autonomy. Friedman explicitly states that agents subject to oppression or subservience can be considered procedurally autonomous on her account, but she includes a threshold mechanism to deal with how these conditions of oppression might give rise to varying degrees of autonomy across agents (2003). Before moving on to analysis of her account of threshold autonomy, I will briefly recount her account of self-reflection. Friedman writes that "when an agent chooses or acts in accord with wants or desires that she has self-reflectively endorsed, then she is autonomous" (2003, p. 5). Her account is essentially identical to Christman's, but instead of

coining the process of critical reflection in terms of self-reflection and nonalienation, Friedman talks in terms of "reflective endorsement" and a "wholehearted commitment" to one's preferences and desires (2003, p. 4-5). Wholehearted reflection is similar to nonalienation as it involves the absence of negative reactive attitudes toward the desire that is in question. So, an agent is autonomous *iff*, when engaging in self-reflection, they come to endorse their preferences wholeheartedly.

Friedman introduces a mechanism by which to assess the extent to which agents are autonomous. The mechanism relies on what *kinds* of preferences or values agents endorse when they wholeheartedly commit to the desires that they are reflecting on, and what effect the kind of preference has on their agency as a whole. Friedman argues that agents can be "content-neutrally autonomous so long as their choices in general accorded with and issued from their deeper wants and commitments." (2003, p. 24) So, even if an agent prefers subservient domesticity, their underlying commitment to values of these kind ensures they are considered autonomous. However, she goes on to suggest that endorsing "traditional femininity" might "diminish the degree of autonomy" the agent exhibits (2003, p. 24). They "still cross the autonomy threshold," but it seems they aren't as autonomous as an agent who endorses non-subservient preferences (2003, p. 24). I think Freidman is attempting to demonstrate that some values or desires might act against an agent's autonomy in the future. That is, endorsing subservience autonomously yields greater possibilities of coercion or manipulation in the future, which is a potential obstacle to being considered autonomous at a later date.

Some might respond to this and suggest that Freidman, instead of remaining content-neutral, is using a substantive conception of the good life in to her measure of degree of autonomy. I think that Freidman avoids integrating a substantive conception of the good in to her account by not making a blanket claim that all instances of subservience will yield less autonomy in the future (2003, p. 20-5). She suggests that some conditions are more likely to do this (2003, p. 24), but does not presuppose they rule out autonomy in virtue of their content. So, instead of saying that the desires endorsed are objectively bad and therefore render the agent non-autonomous (a substantive account), she highlights the way some preferences might decrease an agent's ability to be autonomous in the future because endorsement of their own oppression can result in coercion, which may restrict agents from self-reflecting and wholeheartedly endorsing preferences in the future.

This brings to light another consideration to be made in self-reflective accounts: are agents who exhibit these varied levels of autonomy autonomous with respect to their particular desire that they are reflecting upon, or are they autonomous more generally as an agent? Meyers frames

this in terms of the difference between episodic and programmatic autonomy (1987, p. 625). Episodic autonomy is when an agent is autonomous with respect to a particular situation, whereas programmatic autonomy has to do with agent's capacities to think about their lives more generally and resolve issues or answer questions that are required of them to achieve their overall goals. For instance, programmatic autonomy might have to do with agent's capacity to answer questions like "where do I want my life to go?", "how can I make this happen?" and "should I do X if it gets me closer to my life goals?". In light of the analysis of autonomy diminishing values we saw from Freidman, it seems like Meyer's account is correct in spelling out how this weakened autonomy would operate. For instance, a woman's decision to stay subservient to her husband does not, by hypothesis, restrict her capacity to be autonomous; she might be episodically autonomous with respect to this single issue, provided self-reflection results in her endorsement of the preference. However, her endorsement of these preferences might restrict her ability to be always episodically autonomous in the future due to the nature of subservience.[10] Even stronger is the claim that it might preclude programmatic autonomy from developing as the big picture questions associated with programmatic autonomy are often hindered when agents are in subservient situations. That is, the subservient woman is likely to be subject to her husband's desires with regard to her life plan – perhaps she is only allowed to pursue certain actions that align with the traditional stereotype of a woman. This means she is subject to the will of her husband and, not only is she likely to be subject to coercion and manipulation that might make episodic self-reflection more difficult, but places her under the direct control of another's will, which fundamentally alters her ability to be programmatically autonomous.

> So far I have argued for an account of autonomy that states the following:
> An agent is autonomous, *iff*, upon self-reflection, they endorse their preferences wholeheartedly and without alienation. That is, they lack negative reactive attitudes when considering the development of their preference and the preference's role in their future. This self-reflection can be in regard to particular desires or wants under certain conditions (episodic), or more generally about their life plan and goals (programmatic).

The final concern left for this procedural account of autonomy is a competency requirement that agents must meet in order to count as reflectively endorsing their preferences in

[10] This seems very close to Elster's account of pre-commitment that I discussed in Chapter 1. Although the rubric is fundamentally the same, insofar as the agent is autonomously preferring something that might later cause a lack of autonomy, Elster's account would deem this initial preference non-autonomous because of its content. It would be an attempt at pre-commitment, undermined by the agent's initial lack of autonomy.

the right way. To demonstrate how this concern operates and why a competency condition needs to be discussed, take the following example of an agent: VINCENT wholeheartedly endorsed his preference to become a lawyer. However, during this self-reflection, he suffered a number of delusions that led him to believe he had undertaken a university degree in law and had a successful part time job as a paralegal. These beliefs about his life were false but still formed the basis for his wholehearted endorsement. That is, had he not believed he was studying law and undertaking a successful paralegal position, he would not wholeheartedly endorse his preference to become a lawyer. It seems like these delusions and incoherence should pose a problem for his status as an autonomous agent. Another similarly related example, or criticism of no competency requirement for self-reflection, comes from Khader, who thinks that some accounts of self-reflection seem to imply that in order to endorse something, an agent must have a requisite level of education (2011, p. 82-3). She thinks this must be incorrect because there are instances where educated agents don't seem to be reflectively endorsing their preferences but they are educated enough to be considered competent; there are also agents who are illiterate with no formal education who seem to be wholeheartedly endorsing their preferences regularly (2011, p. 82). So, what we see here is an example of the two potential extremes and subsequent failings of an unrestricted self-reflexivity principle. It can either be too strong so as to require too much of the agent, or it can allow for complete internal incoherence by agents, both of which fail to meet intuitions about autonomy.

To combat these concerns, I propose appealing to the various competencies set out by Christman, Freidman and Meyer. Christman suggests a basic, minimal competency condition of agents' capacities to form an intention to act. These intentions must have no clear "manifest contradictions" (2009, p. 155). This requirement seems to suggest that coherence, rationality and the capacity to intend to act are required for agents to be considered competent. Christman makes a note this competency does not extend to agent's carrying out their intentional actions, as he recognises some things that are out of agents' control might prevent their capacity to act upon their intentions (2009, p. 154). Fundamentally, Christman seems to want to avoid situations such as Vincent's giving rise to autonomy, so he poses that some basic rationality requirements be put in place to ensure the self-reflection nonalienation principle is not subject to objection on these grounds. I think this consideration is sound and basic to all theories of autonomy.

Friedman and Meyer both endorse a view of self-reflection that requires a bit more of the agent, but not so much that agent's belonging to particular groups (for instance, illiterate agents or agents with low intellectual ability) are automatically excluded from the possibility of endorsement. Meyers takes the position that competency sometimes includes exercising imagination, reasons and volitional skill (p. 20-1). Freidman similarly writes that the agent "need not be conscious" of

their endorsement, nor does the reflection have to be "deliberate or deliberated" (2003, p. 8). Additionally, on Friedman's account, reasons needn't take the form of hard evidentiary facts, but instead can take the form of "emotions and desires" (2003, p. 10). So, there seems to be an implicit rejection in this account of extremely cognitive or highly rational requirements making up the self-reflection principle. To demonstrate this point, an agent need not have all the available evidence and use it with awareness to make a judgment to justify their endorsement of their preference. Instead, they can have a positive affective response to their preference during a daydream about the preference's inclusion in their future – imagine an agent who desires to be a housewife; they are having a daydream about the prospect of maintaining a household and being a full time mother and are reacting positively to this daydream. This emotional response is not conscious per se, nor actively a form of self-reflection, but still amounts to reflective endorsement. Alternatively, a negative emotional response to the daydream might cause the preference to be not autonomous, as it is an alienated response to the desire. So, emerging from these accounts is a minimum competency condition that seeks to resolve the issue of self-reflection being construed of as too difficult or too easy to attain.

In light of this discussion, I add the following elements to my procedural account of autonomy: firstly, the agent must be minimally rational. That is, they exhibit basic internal coherence and have the capacity to form intentions to act; secondly, self-reflection can take the form of imaginative, subconscious or unaware endorsement of their preferences, either through the previously discussed forms of nonalienation or wholeheartedness.

3.2. *Are adaptive preferences autonomous?*

In the cases of Jayamma and Jo explored in Chapter 1, there are no prima facie reasons to think that their preferences for wage discrimination and domestic abuse are, by hypothesis, not autonomous. This is because both of their preferences can, hypothetically, be reflectively endorsed under the model of procedural autonomy I argued for in the previous section. Since I moved from from a substantive to a procedural account of autonomy, there can be no reliance on the object of the agent's desire in discerning whether the agent is autonomous. The fact that Jayamma and Jo endorse what might be taken to be objectively bad ends, such as wage discrimination and abuse, does not imply that they are not autonomous because the content of the preferences does not constitute autonomy undermining conditions. Additionally, simply because agents can be autonomous with respect to their adapted preference, does not render the preference no longer adapted, as I will go on to demonstrate.

An initial concern for my claim is that during the course of self-reflective endorsement by the agent, the causal process of the AP is sometimes made apparent to them. Therefore, while the preference was caused by the agent's limitations in options behind their back, the agent becomes aware of the fact that the limitations on her options caused her to form the preference. This would vitiate the necessary conditions of APF that I set out in Chapter 1. So, although the agent may be autonomous with respect to that preference, they are no longer holding an AP. An example of this might take the following form: if Jayamma were to reflect, she would have to engage with questions about why her preference is what it is, which could lead her to discover the 'true' cause of her preferences – viz., that they are the result of limitation on her options. Similarly, Nussbaum recounts that one of the Indian women she interviews, Vasanti, later became aware of the adaption of her preference to stay in an abusive relationship, and that recognising this causal history spurred a repudiation of the preference (2001, p. 68). This is a fair criticism against my account, and it might be the case that some reflective endorsement does bring the causal story to 'the front' of the agent, but this criticism doesn't show that all instances of reflective endorsement render the preferences no longer the result of limited options. My characterisation of self-reflective endorsement is compatible with the causal process of an agent's preferences remaining behind their back.

The above criticism presupposes an incredibly high standard of self-reflective endorsement. Consider how difficult it is to be aware of the actual casual story of regular type preferences – even in regular preference formation, it is almost impossible to say, with certainty, what exactly caused my preference. Perhaps my reasons for undertaking a year of honours study will remain unclear to me forever, but simply because it is difficult for me to get to the bottom of the causal trajectory of this preference does not mean that I'm not autonomous with respect to the preference. Moreover, my account of procedural autonomy explained in the previous section does not require self-reflection to aim for this, nor need it take this form to count as a condition for autonomy. My approach is aligned with the Friedman-Meyers anti-hyper-rationality approach to self-reflection. Instead of focussing specifically on uncovering the causal processes that highlight the reasons for a preference, the endorsement during self-reflection can be grounded in affect and the imaginary. For instance, Jo may wholeheartedly endorse her situation through positive emotional responses to considering the reasons for remaining in a relationship with her abuser, or she might feel nonalienated from an imagined future that continues in a similar vein. To some this might seem to be a shocking conclusion to accept. It's at this point that I bite the bullet and urge the reader to revisit my arguments in favour of accepting a procedural account.

One remedy for those unwilling to accept that agents with AP can be autonomous is the distinction that can be made between episodic and programmatic autonomy in cases of APF. For instance, provided Jo reflectively endorses her preferences, she is autonomous with regards to it. This is an instance of episodic autonomy as it is in relation only to one particular preference, and not an issue that pertains to the agent as a whole. It might be the case that endorsing abuse from a partner (even when it is episodically autonomous) prevents programmatic autonomy. Because it may result in an overall inability to be self-governing or achieve one's big-picture desires. The autonomous preference for domestic abuse prevents Jo from pursuing her life-plan goals or desires such as maintaining friendships or pursuing a career, which prevents her from being programmatically autonomous. This distinction is important in resolving the discontent with calling agents with AP equally as autonomous, or autonomous in the same way as agents who don't hold AP. It demonstrates that although one might be procedurally episodically autonomous with regard to their AP, this preference might violate their capacity to be otherwise programmatically autonomous. This needn't mean that AP always prevents programmatic autonomy, but identifying that it sometimes might goes some way in resolving this tension that remains for the unconvinced.

3.3. *Concluding thoughts on respect and intervention*

The persisting problem for considering agents' AP autonomous is that we lose grounds on which to override their preferences. Any philosopher not committed to a culturally relativist account of morality is likely to want to take this approach if they identify the content of AP or conditions under which they are formed as being objectively morally bad. Moreover, it seems that a motivation for calling APF substantively non-autonomous is to maintain justifiable grounds on which to deny the fulfilment of these preferences or justify social and political reform that attempts to rectify the 'problem' with APF (Khader 2011, p. 74-80). I think that with this underlying consideration in mind, the philosophers discussing APF have focussed on the product of problematic conditions (that is, agents AP), and constructed an indirect theory of autonomy around it to justify interfering with these preferences. As I have discussed in Chapter 2, this approach to autonomy fails in descriptive accuracy and has problematic normative implications. Since I have provided good reasons not to accept a substantive picture of autonomy and henceforth shown that APF is not, ex hypothesi, non-autonomous, I will consider alternate routes that can be appealed to in resolving the persisting problem that APF warrants some kind of intervention.

It's at this point that I explicitly bite another bullet (hopefully a small and palatable one) and suggest that autonomous AP should not be overridden or interfered since they are prima facie autonomous. Autonomous preferences and actions demand respect both interpersonally and politically. On almost any procedural account of autonomy, respect is owed to agents' desires and actions *because* of their autonomy. A motivating reason for the study of autonomy is because we want a way to discern between agent's worthy of respect and those we can justifiably override. If we can accurately discern what we mean when we say "an agent is autonomous", then a theory should allow us to identify autonomous agents and treat them accordingly. In turn, the only way to undermine this commitment to respect of agent's preferences is to either say: (1) that they are not autonomous; (2) argue that autonomy is only one value among many and that there are sometimes cases when alternate values override the need to respect autonomy; or (3) say that autonomy demands no respect at all. I will concern myself with (2), because I have argued against (1) for the most part of this thesis and take option (3) to be at odds with the fundamental commitments of my project.

An example of the second kind of objection is when the state seeks to intervene in somebody's autonomous preferences to ensure its citizen's safety or prevent harm (Sher 1997, 45-100). Examples of when competing values override autonomy include autonomous psychopaths with plans to commit murder or autonomous white supremacists with a desire to commit hate crimes. Here, despite autonomy, many philosophers would make appeals to the need to reduce potential harm done to others (a consequentialist type appeal; Pettit 1989) or to competing values of justice and safety (a liberal appeal; Hurka 1993) to justify interference with the agent's preference. However, taking this approach to cases of APF is not clearly justified, as there is not always a direct harm or threat to other agents simply because some hold adapted preferences. Even appealing to other values overriding considerations of autonomy are met with hostility when no clear demonstrable harm is imminent (Rawls 1993 and Kymlicka 1989). For example, Jayamma's preference for unequal wages is not at all effectively similar to the psychopath's desire to commit murder. Her preference has no direct harm on others, nor does it constitute a complete undermining of commitments to justice in any material way. These cases of autonomous AP are fundamentally characteristically different than the psychopath's and should be treated as such.

So, instead of seeking to override the preferences agents autonomously hold, I suggest a broader attempt to increase an agent's options through attempting to override the oppressive conditions under which these preferences are formed.[11] This approach ensures respect for

[11] This means that I'm committed to letting autonomous AP that are not formed under oppressive conditions, and whose content is not the kind that supports the perpetuation of oppression or discrimination, continue without interference or overriding. After all, the fox's desire to eat something other than the grapes is not motivating the

autonomous preferences, despite their content and formation process, while also tackling the problem of problematically restricted option that underpins APF. The scope of this thesis prevents me from fleshing out the mechanisms by which one could effectively seek to 'fix' oppression, but I hope to shed some light on reasons for taking this approach and suggest how my approach would resolve APF.

The primary concern with my approach is that intervening in oppressive conditions successfully would necessarily involve overriding AP formed under these conditions. My opponents might suggest that in order for oppression to be properly eradicated, the preferences themselves must be overridden. However, it's not clear that this must always happen to successfully change oppressive conditions. I think that implementing policy to enforce institutional change has the capacity to stop APF by dismantling oppressive conditions. I concede that this is an idealistic approach, but the permanent solution to APF should seek to eradicate the cause of APF, not allow APF to form then proceed to stop agents from fulfilling their preferences. If some kinds of oppression restrict options literally (through restricting access to clean drinking water, or the ability to flee violence) or imaginatively (through stereotyping and the social imaginary), then getting rid of the oppression itself should give agents a vaster range of options from which to choose. If the options of oppressed people were broadened then, in theory, there would start to be a decrease in agents holding AP. After all, if what causes AP is a restriction of options, then what would stop the formation of AP is increased options.

This approach does not necessarily result in no agents holding preferences for violence or wage discrimination. Moreover, it might be the case these kinds of preferences are not only held by an agent, but that they include a commitment to enforce these preferences on future generations or their family. For instance, if Jayamma continues to hold her preference for wage discrimination, despite institutional change that seeks to eradicate unequal wages, and teaches her daughters to hold these preferences, or actively restricts their options so that they are incapable of earning as much as their male counterparts, then perhaps we do have grounds on which to interfere with Jayamma's actions. In cases such as these, I think that an agent's preferences are perhaps not in and of themselves unjustifiable and therefore attracting criticism, but instead constitute an interference with attempts to eradicate oppressive conditions and broaden an agent's options. So, overriding them would constitute an attempt to eradicate oppressive conditions, even when the agent holds the preference autonomously.

concern that we ought to do something about APF; the cases of interest are those formed under seemingly oppressive conditions.

Historically, the approach of simply overriding autonomous AP without addressing institutional problems has not only undermined the importance of respecting autonomous preferences, but has also failed to make meaningful change to the conditions that produce AP. It seems that beyond stopping the exercising of some adapted preferences, we want agents to change their minds about what they prefer by fostering conditions in which they have access to a broader range of options. Failures of this kind can be seen in "no-drop" policies for domestic violence charges. Under this policy, primarily practiced throughout America,[12] prosecutors are able to override a victim's preference to withdraw their complaint of domestic and continue with the prosecution of the accused. Often this not only involves overriding the victim's will to drop the case, but in some instances goes so far as to force them to testify against the accused (Davis 2001). Policies like this are attempts to deal with problems of domestic violence by overriding potentially autonomous preferences held by victims of domestic abuse.[13] The failure of this policy to address the broad structural problems that *cause* the victim's preference to drop charges is demonstrated in the fact that there has been no general decrease in the level of domestic violence where this policy is active and no decreased recidivism rates of those successfully prosecuted (Davis 2001; Vincent 2015). Even if the aim of the policy was simply to ensure the victim's safety, there is no empirical evidence to show that this was achieved (Vincent 2015). Therefore, not only has unwarranted interference with autonomous AP constituted a violation of the kind of respect autonomy demands, it has also failed to achieve the overarching goal of reducing rates of domestic violence. In light of such failures, the best way to maintain respect for autonomy and resolve the problem of APF is to deal with the *cause* of the restriction in options that ultimately results in the AP of victims to drop the cases against their partners. Attempts to dismantle domestic violence culture – and in turn, decrease incidences of domestic violence and overcome difficulties in prosecuting offenders – should be actioned through structural change to institutional mechanisms that have direct bearing on the oppressive conditions under which domestic abuse arises. For instance, an increase in accessible women's shelters would directly effect the options victims have after and during their abuser's prosecution. Additionally, law reform that changes the ways in which victims testify during trial to ensure the experience is less traumatising and more supportive would work to make going forward with a prosecution a more viable option for victims.

However, developing a mechanism for this effective and justified institutional change is exceptionally difficult – even the most revered philosophers have admitted that fleshing out the

[12] Studies show that of 142 large prosecutors' offices in the United States, 66% have adopted no-drop policies; (Vincent 2014).

[13] It is not always the case that the victim's preference to drop the case is a *result* of APF, nor is it always the case that they're autonomous, but it is conceivable that at least some instances of agent's wanting to withdraw charges against their accuser is a preference formed due to adapted preferences.

kinds of institutional change necessary would take countless books (Anderson 2012). So more broadly speaking, the aim should be to dismantle oppressive power structures in society and increase members of oppressed groups' options. Specific institutional based changes of this kind include efforts to institutionalise affirmative action in hiring and structural changes that demonstrate a commitment to humane refugee processing policies. These kinds of institutional reforms are the things that give oppressed agents options: affirmative action means that minorities in the workplace are better represented and broadens the potential options available to members of minority groups entering the workplace, and better refugee policies gives oppressed people the option to flee political or religious persecution and violence. The removal of barriers should work to increase the options agents have more generally and result in less APF.

Although these are imprecise recommendations, my hope is that my argument against overriding autonomous AP might justify a discussion about these kinds of potential solutions, and steer us away from the search to find grounds on which to override autonomous preferences. Most importantly, I hope to show that an approach of institutional reform resolves the persisting competing intuitions that stirs inside us when we delve into APF territory: that is, that there is something fundamentally problematic about cases of AP that needs addressing, but there remains a concern that we ought to respect agents' preferences despite their content and their formation process. Seeking to resolve the root of APF promises to prevent the formation of AP under conditions of oppression altogether, while ensuring respect for autonomy is preserved.

Conclusion

I began this thesis with an explicit articulation of the process of adaptive preference formation by interpreting and analysing the works of Elster and Nussbaum, which despite their brilliance, tended to vaguely engage with the concept without precisely pinning it down. I did this work in part because I consider clear articulation of phenomena a necessary pre-cursor to discussions about their importance in moral and political debates, but also in part because having a clear understanding of what adaptive preference formation is allows for a more thorough exploration of whether the preferences themselves are problematic. In turn, I considered how the question of whether adaptive preferences are autonomous is essential to discerning whether they are problematic and justifiably overridden. This discussion of autonomy constituted the bulk of my thesis and involved not only demonstrating that adaptive preferences should not be understood as autonomy deficits, but that the substantive account of autonomy used to justify this view is wrong. Alternatively, I argued for a procedural account of autonomy that I take to be descriptively accurate and normatively useful. In light of my account, I found that adapted preferences are prima facie autonomous, despite their content and means of formation. This conclusion led me to consider how we ought to resolve the intuition that autonomy affords certain respect, while also acknowledging that there remains something unnerving about the conditions under which many adapted preferences are formed.

I see my discussion of adaptive preference formation as forming a small part of a larger problem in moral and political philosophy: balancing respect for autonomy with eradicating unwarranted oppression. That is, finding a way to navigate the persistent intuition that we must resolve unfair and unjustifiable conditions of oppression while also maintaining respect for an agent's autonomy. Striking this balance becomes increasingly difficult when autonomous agents hold a preference for their own oppression, and this is what makes the literature of adaptive preference formation so engaging. In conclusion, I hope this thesis has raised more questions than it has answered. In particular, I hope that it has created a necessary tension between striving to fix the kinds of conditions that restrict people's options, while also demonstrating that preferences formed under these conditions are not, by hypothesis, not autonomous.

Arneson, Richard (1991). "Autonomy and Preference Formation," in Jules Coleman and Allen Buchanan, eds. *In Harm's Way: Essays in Honor of Joel Feinberg*, Cambridge: Cambridge University Press, pp. 42–73.

Benson, Paul (2005). "Feminist Intuitions and the Normative Substance of Autonomy," in J.S. Taylor (ed.) (2005), pp. 124–42.

Berlin, Isaiah (1969). "Two Concepts of Liberty," in *Four Essays on Liberty*, London: Oxford University Press, pp. 118–72.

Berofsky, Bernard (1995). *Liberation from Self*, New York: Cambridge

Burton, Neel M.D. (2012). *Hide and Seek: Understanding self-deception, self-sabotage and more,* Acheron Press.

Christman, John (1990). "Autonomy and Personal History", *Canadian Journal of Philosophy*, 20: 1–24.

—— (1991). "Liberalism and Individual Positive Freedom", *Ethics*, 101: 343–359.

—— (2004). "Relational Autonomy, Liberal Individualism, and the Social Constitution of Selves", *Philosophical Studies*, 117: 143–164.

—— (2009). *The Politics of Persons. Individual Autonomy and Socio-historical Selves*, Cambridge: Cambridge University Press.

Davis, Robert; Smith, Barbara; Davies, Heather (2001). "Effects of No-Drop Prosecution of Domestic Violence Upon Conviction Rates", *Justice Research and Policy*, 3: 1–13

Dworkin, Gerald (1988). *The Theory and Practice of Autonomy*, New York: Cambridge University Press.

Elster, Jon (1983). *Sour Grapes: Studies in the Subversion of Rationality*. Editions De La Maison des Sciences De L'Homme.

Fricker, Miranda (2007). *Epistemic Injustice: Power and the Ethics of Knowing,* Oxford: Oxford University Press.

Govier, Trudy (1993). "Self-Trust, Autonomy and Self-Esteem," *Hypatia*, 8 (1): 99-119.

Harman, Graham (1996). "Moral Relativism," in G. Harman and J.J. Thompson (eds.) *Moral Relativism and Moral Objectivity*, Cambridge MA: Blackwell Publishers, 3–64.

Hurka, Thomas (1993). *Perfectionism*, New York: Oxford University Press.

Kant, Immanuel (1785). *Grounding for the Metaphysics of Morals*, in I. Kant, *Ethical Philosophy*, James W. Ellington, trans., Indianapolis, IA: Hackett Publishing Co 1983.

Khader, Serene (2009). "Adaptive Preferences and Procedural Autonomy", *Journal of Human Development and Capabilities*, 10: 169–187.

—— (2011). *Adaptive Preferences and Women's Empowerment*, New York: Oxford University Press.

Kymlicka, William (1989). *Liberalism, Community and Culture*, Oxford: Clarendon.

Mackenzie, Catriona, and Stoljar, Natalie, eds. (2000a). *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self*, New York: Oxford University Press.

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

—— (1989). *Self, Society, and Personal Choice*, New York: Columbia University Press.

—— (1994). *Subjection and Subjectivity: Psychoanalytic Feminism and Moral Philosophy*, New York: Routledge.

—— ed. (1997). *Feminist Rethink the Self*, Boulder, CO: Westview Press. —— (2004). *Being Yourself: Essays on Identity, Action, and Social Life*, Lanham, MD: Rowman and Littlefield.

Nussbaum, Martha (1997). "Flawed foundations: the philosophical critique of (a certain type of) economics", University of Chicago Law Review, 64:1197–14

—— (2000). Women and Human Development: The Capabilities Approach. Cambridge University Press

—— (2001). Symposium on Amartya Sen's philosophy: 5 Adaptive preferences and women's options. *Economics and Philosophy, 17*(1), 67-88

Prinz, J.J. (2007), *The Emotional Construction of Morals*, New York: Oxford University Press.

Railton, Peter (1986). "Facts and Values," *Philosophical Topics,* 14(2):5-31.

Rawls, John (1971). *A Theory of Justice*, Revised edition (1999) Cambridge, MA: Harvard University Press.

Richardson, Henry (2001). "Autonomy's Many Normative Presuppositions", *American Philosophical Quarterly*, 38: 287–303.

Ripstein, Arthur (1999). *Equality, Responsibility, and the Law*, Cambridge University Press.

Sankey, Howard (2010). "Witchcraft, Relativism and the Problem of the Criterion", *Erkenntnis*, 72(1): 1–16.

—— (2011). "Epistemic Relativism and the Problem of the Criterion", *Studies in History and Philosophy of Science Part A*, 42(4): 562–570.

Sher, George (1997). *Beyond Neutrality: Perfectionism and Politics*, Cambridge: Cambridge University Press.

Terlazzo, Rosa (2016). Conceptualizing Adaptive Preferences Respectfully: An Indirectly Substantive Account. *Journal of Political Philosophy* 24 (2):206-226.

Vincent, Jolene (2014). *Domestic Violence & No-Drop Policies: Doing More Harm Than Good?,* University of Central Florida.

Wolff, Robert Paul (1970). *In Defense of Anarchism*, New York: Harper & Row.