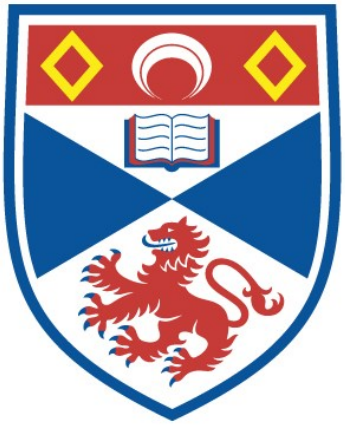
The liberal case for transformative manipulation

Colin Alexander McLean

A thesis submitted for the degree of PhD
at the
University of St Andrews



2024

Candidate's declaration

I, Colin Alexander McLean, do hereby certify that this thesis, submitted for the degree of PhD, which is approximately 55,000 words in length, has been written by me, and that it is the record of work carried out by me, or principally by myself in collaboration with others as acknowledged, and that it has not been submitted in any previous application for any degree. I confirm that any appendices included in my thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

I was admitted as a research student at the University of St Andrews in September 2018.

I confirm that no funding was received for this work.

Date 25/03/24 Signature of candidate

Supervisor's declaration

I hereby certify that the candidate has fulfilled the conditions of the Resolution and Regulations appropriate for the degree of PhD in the University of St Andrews and that the candidate is qualified to submit this thesis in application for that degree. I confirm that any appendices included in the thesis contain only material permitted by the 'Assessment of Postgraduate Research Students' policy.

Date 25/03/24 Signature of supervisor

Permission for publication

In submitting this thesis to the University of St Andrews we understand that we are giving permission for it to be made available for use in accordance with the regulations of the University Library for the time being in force, subject to any copyright vested in the work not being affected thereby. We also understand, unless exempt by an award of an embargo as requested below, that the title and the abstract will be published, and that a copy of the work may be made and supplied to any bona fide library or research worker, that this thesis will be electronically accessible for personal or research use and that the library has the right to migrate this thesis into new electronic forms as required to ensure continued access to the thesis.

I, Colin Alexander McLean, confirm that my thesis does not contain any third-party material that requires copyright clearance.

The following is an agreed request by candidate and supervisor regarding the publication of this thesis:

Printed copy

No embargo on print copy.

Electronic copy

No embargo on electronic copy.

Date 25/03/24 Signature of candidate

Date 25/03/24 Signature of supervisor

Underpinning Research Data or Digital Outputs

Candidate's declaration

I, Colin Alexander McLean, hereby certify that no requirements to deposit original research data or digital outputs apply to this thesis and that, where appropriate, secondary data used have been referenced in the full text of my thesis.

Date 25/03/24 Signature of candidate

Acknowledgements

Neither the commencement nor completion of this thesis would have been possible without the support of many people over the years. First and foremost are the members of my family, and in particular my mother Marian Rankin to whom I owe a debt of gratitude that is beyond measure. My thanks to Anju Puri, Clotilde Torregrossa, Joseph Bowen, Libby South Holland, Pawel Borowski, and Quentin Pharr for their invaluable friendship during my time in St Andrews. I have greatly benefited from the supervision of Dr. Benjamin Sachs-Cobbe, whose feedback and encouragement have played a major role in the development of this thesis, and the incisive comments and suggestions of Dr. Fay Niker. I would also like to thank Dr. Richard Lipsey and Dr. Nancy Olewiler, both of whom contributed to my decision to pursue postgraduate work in philosophy. And finally, my thanks to Dr. James Dean for a cherished but all too brief friendship, and Dr. Kirstie Laird for her encouragement so many years ago. This thesis is dedicated in part to their memories.

Abstract

The liberal philosophical tradition is defined in part by a commitment to political conditions that reflect a view of persons as independent agents who are capable of determining for themselves what matters in life. Intuitively, this requirement places strong restrictions, or even a prohibition, on public policies that aim to affect a change to the target's normative commitments as a means of achieving specific policy goals (transformative policy). This thesis examines a particularly objectionable kind of transformative policy, namely, one that utilizes manipulation to affect the desired change (transformative manipulation). I argue, first, that the strongest case for an absolute prohibition on the use of transformative manipulation is one based on a principle of respect according to which the unconditional value of persons *qua* persons is realized in part by their being reasonably able to exercise a basic kind of autonomy; second, that this principle of respect in fact justifies the use of transformative manipulation when it is necessary to address threats to the stability of liberal political conditions; and third, that we can identify plausible cases where individuals pose a threat to stability that satisfies this condition. I conclude that the liberal tradition can accommodate the use of transformative manipulation.

Contents

Introduction

What are the limits of public policy within the liberal state *qua* liberal state? We might interpret this as a question about the legitimate aims of state action, or alternatively, the means by which the state is permitted to pursue its legitimate aims. This thesis is concerned with the latter question. In particular, I am interested in whether liberal states are permitted to manipulate a target population with the explicit purpose of affecting a change to their normative commitments - e.g. preferences, values, etc - so as to achieve its policy goals. What makes this an interesting question is the apparent unanimity amongst liberal theorists on the in-principle impermissibility of this kind of state action. Unanimity on such matters is unusual. Liberalism is not a singular doctrine, but rather an agglomeration of more or less loosely-related philosophical, political, and economic doctrines that intersects with rival traditions in various ways. It is therefore exceedingly difficult to specify a set of commitments that all and only self-professed liberals endorse. As Stephen Wall points out

> The swathe of ideas it [i.e. the liberal tradition] covers is so broad...that efforts to identify its essential and distinctive features almost always come off as hopelessly narrow...Rather than identifying a single unifying commitment, others have sought...to pick out family resemblance characteristics...True, the more characteristics that are picked out, the less restrictive the resulting characterization of liberalism becomes, but, at the same time...makes it harder to view liberalism as a distinctive tradition of thought, one that differs in deep and informative ways from rival political traditions such as conservatism or republicanism[1]

One way this diversity manifests itself is in disputes about the means by which states are permitted to go about pursuing their (legitimate) aims. All agree that coercion can be justified, but there is no consensus on *when* it

[1] Wall 2015, p. 1.

is justified. Some theorists argue that manipulation is a legitimate part of the policy toolkit while others regard it as antithetical to the liberal ethos. Some argue that the use of the relevant means must be justifiable to their targets for reasons they would accept, while others either deny the principle or have their own interpretation of the apposite reasons. And yet, as we'll see in later chapters, there is something of a tacit consensus amongst liberals that what I am calling 'transformative manipulation' is wholly incompatible with liberalism as each understands it.

This general reticence is undoubtedly rooted in one of the animating themes of the liberal tradition, namely, that political conditions should reflect a conception of persons as independent agents who are capable of determining for themselves what matters in life. As indicated above, theorists cash this claim out in myriad ways. What unites them, however, is the idea that from a political standpoint, there is something sacrosanct about a person's normative commitments. Consider Christman and Anderson's assertion that

> The autonomous citizen acts as a model for the basic interests protected by liberal principles of justice as well as the representative rational agent whose hypothetical or actual choices serve to legitimize those principles[2]

The state can, and must, limit individuals' actions under various circumstances. No one disputes this. But its remit does not extend to shaping what an agent regards as worthwhile or valuable, at least not without their taking an active role in the process. To do so is to fail to treat them in accordance with the conception of a person as described above. The apparent incompatibility of transformative manipulation with the liberal tradition derives from the fact that such measures comprehensively fail to engage with persons *as* persons.

My central claim in this thesis is that despite appearances, transformative manipulation is not incompatible with every plausible conception of liberal political morality. Its use can be justified for the purposes of maintaining the stability of liberal political conditions if there is sufficient reason to believe that alternative policy measures are inadequate. Indeed, I argue that under these circumstances, the use of transformative manipulation is an expression of respect for persons, even members of the target population. My discussion proceeds as follows.

In Chapter 1, I define policy means as state actions that aim to bring persons to regard themselves as having sufficient reason(s) to act in ways

[2] Christman and Anderson 2005, p. 1.

that contribute to the achievement of the relevant policy ends. How they are brought to do so is determined by four constitutive features of policy means: instrument, method, mode, and content. Transformative manipulation is modelled as a particular specification of mode and method. I argue that recent work on the epistemic and moral significance of transformative experience neglects important questions about the political morality of transformative choice for others by reducing it to a matter of interpersonal morality. I then motivate my focus on transformative policy means that utilize manipulative methods specifically by noting that this is the most pernicious form of transformative policy from a liberal perspective. By addressing this 'hard case', we can get a sense of whether there is room within the liberal tradition for less objectionable kinds of transformative policy means.

In Chapter 2 I introduce the Absolute Prohibition Thesis (APT), which states that no plausible conception of liberal political morality can accommodate the state's use of transformative manipulation. The foundation of APT is the liberal presumption that interference with individual freedom is morally wrong unless the interferer can provide adequate justification for their actions with respect to the interferee. This is called the Fundamental Liberal Principle (FLP). Since liberals do not all endorse the same conception of freedom, FLP supports APT iff transformative manipulation conflicts with any plausible conception of liberal freedom, and in a way that cannot overcome the presumption of non-interference on any such conception. Through a descriptive analysis of the concept of freedom, I argue that this is equivalent to the claim that for any plausible conception of liberal political morality, transformative manipulation undermines the satisfaction of conditions that realize the value of agency in a manner that is morally impermissible.

Chapter 3 identifies what I take to be the strongest case for APT. I argue that any plausible conception of liberal political morality endorses certain basic political commitments which entail that an agent's being reasonably able to exercise what I refer to as basic autonomy is a constitutive part of realizing the value of agency - i.e. freedom. Transformative manipulation undermines their ability to do so, and therefore interferes with any plausible conception of liberal freedom. The strongest argument for the in-principle impermissibility of undermining freedom in this way is that doing so necessarily violates a principle of respect for persons that requires the state to engage with them *as* persons. I note, however, that since the state has a respect-based duty to protect the stability of political conditions under which persons are reasonably able to exercise basic autonomy, and discharging this duty sometimes requires treating people in ways that fail to engage with them as persons, then respecting persons doesn't necessarily require engaging with them *as* persons. This argument refutes APT as a conceptual thesis. However, unless

we can identify plausible cases where such treatment is in fact necessary, then APT survives as a *de facto* thesis.

Chapter 4 lays the groundwork for my argument against the *de facto* version of APT. I analyse stability in terms of compliance and enforcement criteria, and highlight how liberal models of stability condition the satisfaction of the latter on the satisfaction of the former. A complete specification of the enforcement criterion determines whether and when the use of different kinds of policy means to address different kinds of threats to stability can be justified. However, extant specifications of the compliance criterion do not provide a sufficiently nuanced picture of the characteristics in virtue of which persons may pose such a threat. Drawing on the concepts of practical and doxastic (un)reasonability, I propose a taxonomy of persons as strongly reasonable, weakly reasonable, weakly unreasonable, or strongly unreasonable, and argue that each category is associated with a unique threat profile.

In Chapter 5 I argue that the *de facto* version of APT fails because there are cases where effectively addressing the threat that strongly unreasonable, weakly unreasonable, and weakly reasonable persons pose to the stability of liberal political conditions plausibly requires the use of transformative manipulation. When persons in each category pose a genuine threat, it is for different reasons. Strongly unreasonable persons pose such a threat because they are unwilling to comply with a liberal order in virtue of normative commitments that reject respect for persons as such. Weakly unreasonable persons are willing to comply with a liberal order for contingent prudential reasons, but become a threat when the group to which they belong becomes sufficiently numerous and/or influential. Weakly reasonable persons endorse respect for persons, but favour non-liberal political conditions that they mistakenly believe are more conducive to basic autonomy. They pose a genuine threat to the stability of a liberal order when they are sufficiently numerous and/or influential enough to control its institutions. I argue that the use of transformative manipulation can be justified in each case, though for different reasons. Since it is in the service of preserving conditions under which persons are reasonably able to exercise basic autonomy, the use of transformative manipulation in these cases is an expression of respect for persons. The *de facto* version of APT must therefore be abandoned.

I conclude in Chapter 6 with some brief remarks on the contribution my account makes to philosophical discussions concerning the limits of public policy in liberal states, and directions for future research.

Chapter 1

Transformative Manipulation and Public Policy

This thesis seeks to address the question of whether transformative manipulation can be a legitimate tool of public policy in liberal states. The present chapter serves two preliminary purposes in connection with this aim. First, it is necessary to clarify the scope of the question. There are many dimensions of public policy, and many ways in which something might be a tool in its service. As we'll see, these dimensions and functions are easily confused or overlooked. To proceed without clearly differentiating them risks conflating a number of conceptual and normative issues discussed in later chapters. The second purpose is a motivational one. That we can ask a question certainly doesn't mean that it's worth devoting significant time to answering. I will argue that the question of whether transformative manipulation is a legitimate tool of public policy has not only been neglected in the relevant literatures, but that there are also at least *prima facie* practical reasons for the state to utilize such measures in certain circumstances. The question therefore has a degree of both philosophical and practical significance that makes it worthy of attention.

My discussion proceeds as follows. In §1.1 I provide a conceptual framework that differentiates between several dimensions of public policy in order to highlight my concern with transformative manipulation as *policy means* specifically. and distinguishes their constitutive features as a means of illustrating how transformative manipulation fits into the general picture. My aim in §1.2 is to demonstrate that current scholarship on transformative experience and choice largely neglects the political morality of these phenomena. After characterizing the idea of a transformative experience, I trace the path from a decision-theoretic puzzle about transformative choice for oneself to

the moral permissibility of transformative choice on behalf of others, and argue that accounts of the latter fail to address certain distinctly political issues that bear on the permissibility of transformative policy means. In §1.3 I argue that the moral status of transformative policy means, though conceptually interesting, also has practical significance. Under certain circumstances, states have at least a *prima facie* reason to utilize such means in the name of efficacy and/or efficiency. Finally, in §1.4 I clarify why I've chosen to focus on the special case of transformative *manipulation*, rather than transformative policy means more generally. The reason is that transformative manipulation is arguably the most objectionable kind of transformative policy from a liberal perspective. Demonstrating that it can be permissible strongly suggests the same of other less objectionable forms of transformative policy means. In this way, my focus on transformative manipulation constitutes a useful first step to a more general account of the political morality of transformative policy.

1.1 Dimensions of Public Policy

Though there is no general agreement on the definition of public policy as such, we can at the most general level distinguish between policy *processes* and policy *options*.[1] The former encompass procedures for identifying policy problems, and formulating, implementing, and evaluating efforts to address these problems. Questions about the legitimacy of policy processes belong to wider debates about what Rawls refers to as the basic structure of society, i.e. "...the way in which the main political and social institutions of society fit together into one system of social cooperation".[2] Though an important topic, I leave the legitimacy of policy processes aside in this thesis. Policy options, on the other hand, specify different ways of addressing policy problems by articulating certain aims (policy ends) as well as how these aims are to be achieved (policy means). Questions about the limits of state action are ultimately about the legitimacy of policy ends and means. They are of three kinds:

1. What kinds of policy ends are consistent with one's political morality?

[1] Smith and Larimer touch on something like this distinction in their claim that "...there is no precise and universal definition of public policy...[Instead] there is a general agreement that public policy includes the process of making choices and the outcomes or actions of particular decisions" (Smith and Larimer 2009, p. 4).

[2] Rawls 2001, p. 10.

2. What sorts of policy means are consistent with one's political morality?

3. Under what circumstances is the *use* of certain policy means permissible?

My inquiry into the permissibility of transformative manipulation concerns (2) and (3). I make no claims about the kinds of policy ends that are consistent with liberal political morality, except that they must be consistent with certain basic political commitments that are discussed in Chapter 3. My more limited aim is to determine whether the use of transformative manipulation as policy means is or is not consistent with liberal political morality, and if it is, the circumstances under which its use can be justified.

Before tackling these questions, some conceptual ground-clearing is in order. In the next section I set out a functional definition of policy means that captures their essential purpose *qua* policy means. In the subsequent section I discuss their constitutive features, the specification of which determines precisely how this function is carried out. These clarifications will both inform the analysis of arguments for and against the permissibility of transformative manipulation in later chapters, and help us to avoid conflating distinct kinds of worries about the substance and use of policy means.

1.1.1 The definition of policy means

Within the policy literature, what I am calling policy means are typically referred to as policy instruments. Verdung, for example, remarks that "Public policy instruments are the set of techniques by which governmental authorities wield their power in attempting to ensure support and effect or prevent social change".[3] As a result, a great deal of discussion about the nature of policy means devolves into the identification of what Dodds aptly describes as a "bewildering number and variety of types of policy instruments", and the creation of a wide variety of classification schemes.[4] For example, Lowi's influential model posits four categories of policy instruments: distributive, regulatory, redistributive, and constitutive.[5] Meanwhile, Dodds proposes a five-category model: use of financial resources, authority, organization, and the provision of information.[6] As we'll see, these frameworks neglect key dimensions of the concept.

My analysis of the permissibility of transformative manipulation is based on the following definition of policy means:

[3] Verdung 2010, p. 21.
[4] Dodds 2013, pp. 23–24.
[5] Lowi 1972.
[6] Dodds 2013, p. 24.

> Actions undertaken or not undertaken by the state (or on behalf of the state) whose aim is to lead individuals to have sufficient subjective reason(s) to ϕ, where their ϕ'ing is judged by policymakers to be necessary and/or sufficient for securing ω

where ϕ is conative or doxastic state or action, and ω is a policy end. This definition has three important features. The first is that it describes policy means as *agent-focused*. Agent-focused policy means target attributes, actions, or states of individuals that are relevant to securing particular policy ends. For example, criminal laws seek to influence behaviour, educational programs to cultivate skills, and health policies to treat/prevent afflictions or mitigate suffering. Now, there are policy means that are not agent-focused. In some cases, achieving the policy ends does not require a change to anyone's behaviour - for example, an initiative to allow spawning salmon to return to an isolated mountain stream by removing obstructions. However, since the vast majority of public policies are designed as direct or indirect solutions to coordination problems, we can leave these exceptions aside for the purposes of the discussion.

Second, my definition characterizes policy means as constitutively *reasons-related*. Reasons-related policy means engage with or otherwise affect individuals' reason (practical or theoretical) in an attempt to bring them to ϕ. For example, laws that prescribe criminal charges for owning a firearm without a license (as part of a policy to reduce firearm-related fatalities, say) provide individuals with reasons to comply with the law. The vast majority of agent-focused policy means are reasons-related in this way. However, there are exceptions. For example, a policy that involuntarily confines severely mentally ill persons who are unresponsive to reasons in order to protect both themselves and others from harm. Or, to take a more sordid case, consider the confinement of American and Canadian citizens of Japanese descent in interment camps during World War Two. The policy means employed, i.e. relocation to internment camps, were not intended to *convince* the target population not to engage in activities that would endanger national security. Rather these means physically prevented them from having any opportunity to do so. Non-reasons-related policy means tend to be the exception, especially in liberal societies. For the sake of simplicity, I therefore leave them aside in my analysis.

The final feature that requires clarification is the idea of a sufficient subjective reason. Reasons can be conceived in two ways: First, as considerations that count in favour of something (normative reasons); and second, as considerations in virtue of which an agent is motivated to do something

(motivating reasons).[7] One and the same reason can be normative and/or motivating. For example, suppose P has objectively good reasons to get up for work on time, but fails to recognize them as good reasons. In this case, P has a normative reason to get up for work on time, but lacks a motivating reason to do so. Conversely, suppose P's addiction compels him to regularly acquire and consume heroin despite his wanting to stop. In this case, P has a motivating reason to acquire and use the drug, but not a normative one. Finally, suppose P is in the early stages of pregnancy. P should take folic acid to prevent certain birth defects, and upon learning this from her physician, she begins to take the necessary supplements. In this case, P has both a normative and motivating reason to take folic acid.

With this distinction in hand, subjective reasons can be understood as considerations that P regards as a normative reason to ϕ, and a *sufficient* subjective reason as considerations that P regards as a normative reason to ϕ of sufficient weight to motivate him to ϕ. The purpose of policy means, then, is to bring individuals in the target population to regard themselves as having sufficiently weighty reasons to act in ways that, either directly or indirectly, contribute to the satisfaction of the relevant policy ends.

1.1.2 Constitutive features of policy means

Policy means can be designed to carry out their function (i.e. lead the targets to regard themselves as having sufficient reason(s) to ϕ) in a variety of ways. They have at least four constitutive features whose specification is an integral part of the design: *instrument*, *method*, *mode*, and *content*. As we'll see, not only are these features logically distinct, they are associated with different normative debates. The reason I make these distinctions is therefore to clarify the boundaries of my analysis of the permissibility of transformative manipulation as policy means.

Instrument

Policy means employ a variety of mechanisms to bring persons to regard themselves as having sufficient subjective reasons to ϕ, what I will refer

[7] Dancy (2000) provides a useful historical overview of the distinction. In line with most contemporary theorists, I regard normative and motivating reasons as two aspects of reasons that highlight how they figure into our rational economy, not two different kinds of reasons (Alvarez 2010; Dancy 2000; Scanlon 1998). I leave aside a third putative category or reasons, namely explanatory reasons, which are commonly defined as considerations that render an agent's actions intelligible.

to as *instruments*. Examples include public expenditures, tax incentives, education programs, advertising campaigns, criminal laws, and regulations.[8] Some theorists have defined instruments by way of more general properties such as visibility, directness, automaticity, resource intensity, precision, and coerciveness.[9] For the purposes of my analysis, we can simply distinguish between coercive and persuasive instruments.[10]

Coercive instruments significantly reduce (or destroy) the appeal of certain options by attaching sufficiently onerous costs to pursuing them.[11] Note that they do not, strictly speaking, take options off the table; rather, they take *combinations* of options off the table, namely, pursuing the option and not incurring the associated costs.[12] Coercive instruments typically take the form of legal penalities such as fines or incarceration, but can also involve offers the refusal of which would involve forgoing a significant benefit.[13] For example, consider a policy to reduce pregnancies amongst chronic opiate addicts by offering them a large cash payment to undergo sterilization. From the addict's perspective, the overwhelming net present benefit of the payment would surely make refusing sterilization almost impossible.

Persuasive instruments, in contrast, are designed to simply communicate reasons for choosing certain options without altering their relative costs. For example, an anti-smoking ad that describes the effects of tobacco consumption on the lungs and heart; or a financial literacy campaign that lays out the risks of taking on large amounts of debt; or mandating product labels that draw consumers' attention to the carbon intensiveness of a product category or industry. In each case, the effect is to provide the target with information about existing costs that is (from the perspective of policymakers) relevant to their deliberations about whether to choose this way or that.

Method

If instruments are the mechanisms by which the state seeks to bring persons to regard themselves as having sufficient subjective reason to ϕ, then the method refers to how the application of instruments is presented to the target.

[8] Cairney 2012, p. 26.

[9] Salamon 1989; Schneider and Ingram 1990.

[10] Theorists commonly represent manipulation as falling somewhere between persuasion and coercion - see Faden, Beauchamp, and King 1986, p. 259 for a clear example. I do not, for reasons that become clear in the next sub-section.

[11] I draw this conception of coercion from Feinberg 1986, Ch.23

[12] Ibid., p. 192.

[13] There is a vast literature on the topic of coercive offers. For some influential accounts, see Zimmerman 1981, Feinberg 1986, and Stevens 1988.

The mark of a *rational* policy method is that an instrument is applied in such a way that the target is aware (or could easily become aware) both *that* it is being applied and *why*. As an example of rational persuasion, consider the financial literacy campaign mentioned above. It could involve making a free set of government-branded learning materials available to the public as well as information campaigns that clearly articulate why the initiative is being undertaken. From the public's perspective, there is no ambiguity about the existence of the initiative, nor its rationale. Or, as an example of rational coercion, consider a policy to reduce rates of drunk-driving by announcing and implementing increased penalities and enforcement actions. The policy is clearly coercive given the potentially life-changing penalities associated with non-compliance. But it is also rational because both its existence and rationale are transparent to the target population.

In contrast, the policy method is *manipulative* when, roughly speaking, the application of an instrument is intentionally obscured or its true rationale is not made clear to the targets.[14] As Raz observes (in perhaps too loaded terms), manipulation "...perverts the way that [a] person reaches decisions, forms preferences, or adopts goals".[15]As an example of manipulative persuasion, suppose the government quietly implements regulations that require banks to both freely provide and aggressively promote their own educational materials on financial literacy to customers. The policy uses a persuasive instrument, since the bank's actions do not alter the costs of financial responsibility or irresponsibility. But it is also manipulative, because the policy means have been expressly designed to be hidden from the target population *qua* policy means. Alternatively, the policy could be manipulative because its apparent rationale is just a means to satisfy undisclosed aims. Suppose that it is advertised as a push for financial literacy because this is something most people would support, but the real motivation is just to get more money into the stock market.[16] Insofar as knowledge of the policy's true aims are relevant to the target population's response to it, the method is manipulative.

Similar modifications can be made to the drunk-driving policy to produce an example of manipulative coercion. Suppose that its stated purpose is to crack down on drunk-driving, since preventing harm to others is a rela-

[14] There are of course numerous ways to think about manipulation. I do not claim that this is the only plausible conception. See Coons and Weber 2014 for an overview of debates on the topic.

[15] Raz 1986, pp. 377–378.

[16] I am assuming that there is some connection between financial literacy and propensity to invest.

tively uncontroversial policy aim. However, in reality, the policy is intended to reduce alcohol consumption for paternalistic reasons. The instrument is clearly coercive for reasons discussed in the previous section. But is also manipulative because the application of the instrument involves withholding information that could be relevant to the target population's deliberations about their support for it.

Mode

The mode of policy means determines how the target's normative commitments are meant to figure into its dynamics.[17] In the *conservative* mode, policy means are only meant to leverage an individual's existing set of normative commitments to lead them to regard themselves as having sufficient subjective reasons to ϕ. Consider, for example, a law (L) to reduce consumption of a substance (S) by prescribing incarceration for possession. Since most people regard incarceration as inimical to their own interests, they will regard L as a strong prudential reason to abstain from S. Note that L is not meant to change anyone's mind about whether they should have a preference for S, or the value of a life that involves consuming S, or even that incarceration is a bad thing. It simply trades on the fact that for reasons of self-interest, most people already have a sufficiently strong aversion to incarceration.

In the *transformative* mode, policy means are designed to affect a change to a person's normative commitments such that they come to regard themselves as having sufficient subjective reason to ϕ.[18] On the one hand, this may involve making them aware of normative commitments that they do not currently endorse but are entailed or strongly supported by their current set. Call these *weakly* transformative policy means. For example, suppose policymakers want to address widespread exploitation of workers in the private sector by strengthening the bargaining power of labour, but face hostile attitudes amongst workers towards unions. In order to accomplish their aims, policymakers begin a campaign to convince workers that they should value unions *because* they already value such things as fair compensation and working conditions. The purpose of the policy means, then, is to alter workers'

[17] By normative commitments, I mean those elements of a person's character that are the sources of subjective reasons, such as their desires, preferences, and values.

[18] Schneider and Ingram's (1990, p. 519) conception of 'symbolic or hortatory' policy tools is quite similar to the concept of transformative policy means; however, like many authors, they conflate what I am calling 'policy modes' and policy instruments/tools, which, based on their different contributions to generating subjective reasons to ϕ, should be distinguished.

normative commitments in a way that is consistent with what they already value.

Transformative policy means can go further than this, however, by affecting a fundamental change to a person's set of normative commitments. In other words, a change that results in an *ex-post* set that is distinct from the *ex-ante* set. Call these *strongly* transformative policy means. As an example, suppose policymakers are required to implement certain climate change mitigation initiatives in the face of public sentiment that is hostile to environmentalist values. In response, they implement a suite of long-term measures that stigmatize such attitudes while at the same time valorizing and rewarding climate change mitigation efforts by implicitly associating them with other values that the population holds, e.g. patriotism, faith, etc. Assuming the efficacy of the policy means, the result would be widespread replacement of anti-environmental values with those that are at least more sympathetic with environmental concerns.

Content

The content of policy means specifies the kinds of reasons that the means are designed to bring persons to regard themselves as having to ϕ. Consider three examples:

(a) A tax incentive for household recycling with the aim of reducing waste

(b) A campaign to reduce drinking and driving by emphasizing the potential for catastrophic harms to others

(c) An initiative to increase public awareness of climate change by providing scientific evidence for its effects

In (a), the policy means are designed to bring persons to regard themselves as having a prudential reason to recycle. No attempt is made to convince them of the rightness of their actions, only that it is in their financial interest. This is not the case in (b). The policy means are designed to bring the target to see themselves as having a moral reason to refrain from drinking and driving, not a merely prudential one. In (c), the provision of clear and accurate information about climate change is meant to provide the target population with epistemic, rather than prudential or moral, reasons. Policy means can, of course, have complex content. Efforts to reduce the incidence of harmful behaviours might appeal to moral considerations while also prescribing legal penalties as prudential reasons for those that are not moved by the former.

1.1.3 Summary

In this section I have highlighted four constitutive features of policy means whose specification determines how a policy is designed to bring persons to regard themselves as having sufficient reason to ϕ (Table 1.1).

Table 1.1: Features and specifications of policy means

Instrument	**Method**	**Mode**	**Content**
Persuasive	Rational	Conservative	Prudential
Coercive	Manipulative	Transformative	Moral
			Epistemic

Most of these specifications have already received significant philosophical attention. Debates about coercion as a tool of state power have of course been central to political philosophy for centuries, and are closely related to questions about the scope of individual liberty. Meanwhile, the nature and permissibility of manipulation by the state has become an increasingly prominent area of research in contemporary political theory, particularly as it relates to the issue of autonomous choice. So too have questions about the kinds of reasons that should be allowed to figure into the justification of state action, which feed directly into debates about state neutrality, perfectionism, and paternalism. In contrast, comparatively little attention has been paid to the nature and permissibility of transformative policy. On the one hand, this shouldn't be surprising. Though the liberal tradition evinces enormous diversity, one of its core features is a commitment to treating individuals as rational agents who are entitled to decide for themselves what matters in life. This standpoint seems to preclude any justification for wielding the considerable power of the state for the express purpose of affecting changes to an individual's normative commitments. And yet, we cannot simply assume that this conclusion is entailed by any plausible conception of liberal political morality. It is a substantive claim that must be demonstrated.

One might object to my claim that philosophers have neglected transformative policies by pointing to the recent explosion of work on the concept of transformative experience. In the next section I would like to argue that although this work is invaluable for clarifying the precise nature of the justificatory challenge posed by transformative policy, current approaches to the justification of transformative choice for others at the level of interpersonal morality cannot be straightforwardly applied at the level of *political* morality. Therefore, even if we assume that one of these approaches is correct, the permissibility of transformative policy is still an open question.

1.2 Transformative Experience

Certain kinds of experiences catalyze fundamentally new ways of feeling, seeing, or knowing, and so cannot be fully understood or appreciated until one has undergone them. They are what theorists have come to refer to as 'transformative experiences'.[19] Consider the case of becoming a parent.[20] Beforehand, one might imagine that the bond between parent and child is just a stronger version of other kinds of bonds, say that with a beloved pet; or one might have very definite opinions about *the* correct approach to parenting. As most parents will attest, not only are these beliefs almost certainly wrong, it is profoundly difficult for someone who has never been a parent (including their earlier selves) to truly understand *why* they are wrong. Any understanding that a non-parent might have could be compared to that of a person who is provided a description of a painting without actually seeing it. Based on what they've been told they might grasp that it must be beautiful, but no amount of description will capture the actual experience of seeing it.[21] Not all transformative experiences are so momentous, however. Those who take up a combat sport frequently report the experience as revelatory in ways that are difficult to convey to anyone who has not undergone the experience. Consider the following reflections by a beginner in Brazilian Jiu Jitsu:

> I can now attest that the experience of grappling with an expert is akin to falling into deep water without knowing how to swim. You will make a furious effort to stay afloat—and you will fail. Once you learn how to swim, however, it becomes difficult to see what the problem is—why can't a drowning man just relax and tread water? The same inscrutable difference...can be found on the mat: To train in [Brazilian Jiu Jitsu] is to continually drown—or, rather, to be drowned, in sudden and ingenious ways—and to be taught, again and again, how to swim.[22]

Clearly, learning Brazilian Jiu Jitsu isn't literally like learning how to swim. The analogy is apt because it captures the experience of struggling to adapt to external forces by developing a specific set of skills. But physically struggling to subdue another agent who seeks to do likewise differs from struggling to

[19] See Ullmann-Margalit 2006 and Paul 2014 for pioneering treatments of the phenomenon.
[20] This example is frequently discussed in the literature on transformative experience. See ibid., pp. 71–94.
[21] Frank Jackson 1982 makes this point in a different context with his Knowledge Argument.
[22] Harris 2012.

adapt to a natural environment in ways that cannot be fully appreciated without first-hand experience with both.

As these examples illustrate, transformative experiences can have both an epistemic and a personal dimension.[23] The former involves the acquisition of knowledge that provides novel ways of conceptualizing or knowing about oneself and the world.[24] To learn how to effectively grapple with an opponent is to be introduced to a domain of knowledge that is quite unlike any other. It involves not just the development of new physical skills, but also knowledge of one's instinctual reactions to physical aggression and mental fortitude under pressure. In contrast, the personal dimension involves a change to one's desires, preferences, or values - in other words, one's normative commitments. Becoming a parent does not merely provide one with cognitive access to new ways of thinking about oneself and the world. For most parents, it also fundamentally alters what they regard as mattering in life, their sense of purpose, their priorities. This is not to say that all transformative experiences involve both dimensions, however. It is perfectly possible to change one's normative perspective without gaining new kinds of knowledge. Similarly, one can come to understand the world in new ways without any alteration to one's normative commitments.[25]

1.2.1 The puzzle of transformative choice

Philosophical interest in transformative experience has centred primarily on a decision-theoretic puzzle about transformative choice.[26] The core the issue is this: rational choice between a set of options involves judgements about their subjective value, i.e. their perceived value. In other words, the choice to ϕ rather than not-ϕ is rational only if we judge that ϕ'ing is the better or more optimal thing to do all things considered.[27] The process by which we make such judgements proceeds on the basis of our current beliefs and normative commitments - what Ullmann-Margalit refers to as our 'rationality base'.[28] Now suppose that person P believes that ϕ'ing will be personally transformative for them, and so fundamentally alter their rationality base.

[23] Paul 2014, pp. 10–16.
[24] See ibid., 11 n.11 where she addresses the worry that this definition might render every new experience a transformative one.
[25] Ibid., p. 17.
[26] See Lambert and Schwenkler 2020a for an excellent collection of essays covering various aspects of the debate.
[27] Note that this is only a necessary condition for rational choice.
[28] Ullmann-Margalit 2006, p. 157.

Can P rationally choose to ϕ? An affirmative answer to this question seems to entail that P can rationally conclude that the *ex-post* rationality base is superior to their current rationality base *on the basis of* their current rationality base. But as Lambert and Schwenkler observe

> If an individual's core values partially determine who they are, then what could possibly determine which set of wholly different values to adopt? More generally, whatever one uses as the basis for all their choices, what about that basis could make it rational to choose a shift to a new basis?[29]

Ullmann-Margalit's solution is to enrich our understanding of rational choice, remarking that it "...need not mean optimizing; it can also mean acting reasonably".[30] Reasonability can take a variety of forms in this context. For example, one might adopt a strategy of incrementalism. Instead of making one big transformative choice, we might break it up into a series of smaller choices each of which is not itself transformative, but whose cumulative result nevertheless is, e.g. instead of making the leap from dating right to marriage, a couple could decide to first move in together, then pool their finances, and so on, until the choice to marry is but a much smaller step. As she notes, however, not all transformative choices can be broken up in this way - in other words "Some cases really call for leaping across an abyss".[31] Here, reasonability might be defined by appeal to higher-order commitments.[32] P's current rationality base might council against joining the military, but doing so may not be irrational if P has a second-order desire to become someone who possesses certain martial virtues. In this way, a non-optimizing decision relative to P's existing rationality base can be reasonable - and therefore rational - even when it comes to irreducibly transformative choices.

Paul agrees with Ullmann-Margalit that many transformative choices that we face over the course of our lives do require leaping across an abyss. However, she (Paul) is not so sanguine about the prospects of resolving the puzzle by appeal to higher-order preferences. One of her worries is conceptual. Suppose that P has a first-order preference $\mathbf{x}$, and also a second-order preference to have a preference for $\mathbf{y}$ instead of $\mathbf{x}$. Surely the fact that P is willing to give up $\mathbf{x}$ in favour of $\mathbf{y}$ means that $\mathbf{x}$ isn't one of P's genuine preferences. As Paul notes (paraphrasing Richard Pettigrew)

[29] Lambert and Schwenkler 2020b, p. 3.
[30] Ullmann-Margalit 2006, p. 168.
[31] Ibid., p. 169.
[32] Ibid., pp. 167–168.

> [I]f you are prepared to dispense with your current preferences in order to take on new preferences, in what sense are your current preferences really your preferences?[33]

If this is correct, then the idea that higher-order preferences can ground rational transformative choice rests on a confusion about what it means to *have* a preference for something.

Paul's second worry is that the appeal to higher-order commitments solves one puzzle but at the cost of generating an equally difficult one.[34] Suppose that P has a second-order preference to have a different set of first-order preferences **y** than their current set **x**. In other words, P_x has a second-order preference to become P_y. This must mean that P_x regards being P_y as in some sense superior to being P_x. But how could P_x make such a judgement given the impossibility of knowing what life as P_y will be like? For all P_x knows, their preferences as P_y will become such that they judge P_x to be superior. Is the rationality of transformative choice in this context defined in terms of P_x's second-order preferences, or P_y's?

1.2.2 Transformative choice for others

In addition to its decision-theoretic significance, the puzzle of transformative choice also has a moral dimension. The difficulty of pinning down the conditions for rational transformative choice for oneself raises questions about the moral permissibility of transformative choice on behalf of others ('vicarious transformative choice').[35] That we make such choices is undeniable. Parents must make decisions on behalf of their children that fundamentally shape their future selves, and families and healthcare professionals are often required to make profoundly consequential choices on behalf of patients. However, such decisions are also made in the political domain. Public policies shape the very structure of social and economic relations between citizens, and so cannot but have epistemically and personally transformative effects. I would like to suggest that current scholarship on the morality of vicarious transformative choice fails to address certain distinctly political concerns that bear on its permissibility as a tool of state action. In other words, there is a gap in the literature when it comes to the *political* morality of vicarious transformative choice.

[33] Paul 2014, 91 n51.
[34] Ibid., pp. 91–94.
[35] I draw this terminology from Howard 2015.

The Adaptive Preferences Test

Dana Howard considers the intuitively plausible claim that vicarious transformative choice can only be justified by appeal to the future preferences or attitudes of the person on who's behalf the choice is made, what she refers to as 'Predictive Glad'.[36] In essence, the idea is that a vicarious transformative choice ϕ on behalf of P is permissible only if we have good reason to believe that P would endorse ϕ *ex-post*. For example, consider a parent's decision to enrol their child in piano lessons in the belief that, despite the child's dislike for it, they will eventually come to endorse the decision as having enriched their life. Despite its plausibility, this kind of justification involves a troubling circularity. As Howard notes

> The tricky part about making decisions over children is that guardians may act not only in ways that they think are in the best interest of the child, but the vicarious decisions that they end up making also shape what the child himself takes to be in his best interest in the long run[37]

In other words, Predictive Glad justifies vicarious transformative choice by appealing to a future rationality base that is itself a product of that choice. In doing so, it effectively begs the question, and so cannot serve as plausible grounds for permissible transformative choice on behalf of others. Does this mean that we can only justify vicarious transformative choice on the basis of P's *ex-ante* rationality base? Clearly not, for this would render many of the most important decisions that parents make on behalf of their child morally suspect. We seem to face a dilemma: accept a question-begging standard of permissible vicarious choice; or implausibly restrict the range of permissible vicarious transformative choices.

Howard's solution employs what I'll refer to as the Adaptive Preferences Test. Preferences are adaptive when they are a subconscious response to restrictions that diminish one's set of feasible options from a larger set of conceivable alternatives.[38] For example, a prisoner might come to enjoy mopping common areas because it is the best of a very limited set of ways of spending time outside of his cell. Howard argues that a vicarious transformative choice ϕ on behalf of P is morally permissible only if

1. We have good reason to believe that P will endorse ϕ *ex-post* on the basis of non-adaptive preferences; and

[36] Howard 2015.
[37] Ibid.
[38] Howard draws this definition from Elster 1983.

2. We have good reason to believe that P would come to regret not-ϕ on the basis of non-adaptive preferences

If (1) is satisfied, then P's *ex-post* preferences aren't a subconscious response to restrictions imposed by ϕ, and so can figure into the justification of ϕ without circularity. As Howard notes, however, it is not enough that P non-adaptively endorse ϕ *ex-post*. Consideration must also be given to whether P would come to non-adaptively endorse or regret not-ϕ being chosen on their behalf. If he would also non-adaptively endorse not-ϕ, then ϕ could not be justified by appeal to his *ex-post* rationality base. Conversely, if he would non-adaptively regret not-ϕ - i.e. if (2) is satisfied - then ϕ is permissible on the basis of his *ex-post* rationality base, for it is the course of action that he would genuinely come to prefer. Howard's concludes that

> We cannot simply appeal to the future pro-attitude of the principal to justify some specific action. Moreover, we cannot appeal to the reasonableness [i.e. non-adaptiveness] of the principal's future pro-attitude alone to justify that action. Instead, we must determine the predicted future pro-attitudes of both courses of action to see if there are any lessons we can draw[39]

Thus, the parent's choice in the piano lessons case would pass the Adaptive Preferences Test if, first, the fact that the child is required to take lessons does not itself preclude him from exploring other options in life, and so developing preferences that may not include playing the piano. If he does endorse the decision *ex-post*, then it is not on the basis of preferences that are adaptive as a result of that choice on his behalf. And second, choosing not to enrol him in piano lessons does not itself preclude experiences that could bring him to regret not having learned to play when he was younger. If he does come to regret this choice having been made for him, then it is not on the basis of preferences that are adaptive in response to this choice either. Insofar as these conditions are satisfied (and all else being equal), the parent can reasonably justify their vicarious choice by appeal to the child's expected future attitudes.

Whatever its merits at the level of interpersonal morality, the Adaptive Preferences Test cannot be easily transposed to the political domain. The fact that *ex-ante* attitudes play no necessary role in the determination of permissible transformative choice on behalf of others means that the test is consistent with a particularly robust form of ends-paternalism that many

[39] Howard 2015.

theorists would reject. As discussed, this isn't a problem at the interpersonal level; if a target's *ex-ante* attitudes were necessarily authoritative then many important and unavoidable decisions in life (e.g. for one's children) would be morally suspect. But the relationship between state and citizen is not strictly analogous to that between parent and child. It is not a given that one of the constitutive roles of states is to make decisions on behalf of citizens about their own good. Indeed, the denial of this claim is a hallmark of the liberal political tradition. Therefore, if we are to apply the Adaptive Preferences Test to vicarious transformative choice by the state, we must first settle certain distinctly political questions about the relationship between state and citizen.

The Global Utilities Approach

In a case of conceptual convergence, nudge theorists often appeal to a justificatory principle that is effectively identical to Predictive Glad. Thaler and Sunstein propose that implementing choice architectures that bias a person's decisions in a specific context in one direction rather than another can be justified if they are made better off as judged by themselves (AJBT).[40] Just as Howard recognizes the circularity challenge that vicarious transformative choice poses for Predictive Glad, so too have these theorists recognized the same for AJBT. Paul and Sunstein, for example, remark that

> Here, then, is the root of the problem for the AJBT criterion. Important choices...can result in endogenous preference change...If our assessment of the value of such changes is merely that, as judged by themselves, people will be happy ex post...this is not sufficient to distinguish between alternatives. If, for each change we consider, as judged by themselves, people will be happy ex post, we need a further criterion[41]

Richard Pettigrew attempts to resolve the problem by way of what I'll refer to as the Global Utilities Approach.[42] He argues that the circularity of AJBT with respect to vicarious transformative choice is attributable to its being defined in terms of local utilities - that is, utilities that an agent assigns to a past, present, or future outcome at a particular time.[43] For example, the utility that P at time t_1 (P_{t1}) assigns to an outcome ω at each of times

[40] Thaler and Sunstein 2008, p. 5.
[41] Paul and Sunstein 2019, p. 7.
[42] Pettigrew 2023.
[43] "These are the utilities that encode their values at that time" (ibid., p. 8).

$[t_{1\text{-}n}...t_1...t_{1+n}]$ are different local utilities of ω for P_{t1}. Justifying a vicarious transformative choice on behalf of P by appeal to AJBT seems to require that we decide whether to take P's *ex-ante* or *ex-post* local utilities as authoritative. As we've seen, Howard defends the authority of *ex-post* attitudes/local utilities under certain conditions. Pettigrew's strategy is to deny that either *ex-ante* or *ex-post* utilities are by themselves authoritative. Rather, to determine whether a vicarious transformative choice to ω on behalf of P_{t1} is permissible, we must calculate the *global utility* (GU) of ω for P_{t1}, which is the sum of the local utilities they assign to ω. Or, more formally:

$$P_{t1}[GU(\omega)] = P_{t1}[(\omega_{t1\text{-}n}) + ... + (\omega_{t1}) + ... + (\omega_{t1+n})]$$

Crucially, each local utility of ω on the right-hand side of the definition is assigned a weight that affects its contribution to the global utility of ω for P_{t1}.[44] In effect, these weights signify how much P_{t1} cares about the local utilities of his past, present, and future self/selves at present. For example, suppose that Bob_{t1} is considering joining a gym to get in shape. Being naturally lazy, he sees little appeal in doing so at present, but expects that *if* he does, he will eventually be very glad that he did. Conversely, his attitude on the matter will not change over time if he doesn't join. Table 1.2. depicts the local utilities that Bob_{t1} assigns to each choice at the present time (t1) and what he expects it will be for his future self (t2). The global utility (GU) for each choice reflect Bob_{t1}'s giving equal consideration to his present and future self. In this scenario, the result is a higher global utility for not joining the gym.

Table 1.2: Equal weights for local utilities

	Bob_{t1}		
Choice	t1	t2	GU
1. Join gym	1	15	16
2. Don't join	10	10	**20**

Now suppose that Bob_{t1} assigns the same local utilities at t1 and t2 for each choice, but also thinks because his future self will derive so much value from his present choice to join (choice 1 at t2), that future self should be given outsize consideration in his deliberations. Table 1.3. depicts this scenario.

[44] The weighting of local utilities is a complex issue. For a detailed discussion see Pettigrew 2023, pp. 9–11.

The local utility that Bob_{t1} presently assigns to joining the gym (choice 1 at t1) has been discounted by 0.5 and his future local utility for having joined (choice 1 at t2) has been multiplied by 1.5 to reflect the greater consideration he gives to that future self in his present deliberations. As a result, his global utility for joining the gym is now higher than not joining.

Table 1.3: Differential weighting of local utilities

	Bob_{t1}		
Choice	t1	t2	GU
1. Join gym	$1_{0.5}$	$15_{1.5}$	**23**
2. Don't join	10	10	20

Pettigrew argues that if a vicarious transformative choice on behalf of P_{t1} to experience ω is permissible, then it must be the case that the global utility of ω is greater than the global utility of not-ω for P_{t1}.[45] Applied to the scenario in table 1.2, this criterion entails the impermissibility of making such a choice on behalf of Bob that will lead him to join the gym. Applied to the scenario in 1.3. however, such a choice is permissible (all else being equal).

But what about cases where P_{t1} has no local utilities, and therefore no global utility, for ω? Is vicarious transformative choice for her to undergo ω permissible? Pettigrew considers two variants of this case.[46] In the first, P_{t1} has no global utility for ω, but if given the opportunity would assign local utilities such that the sum would favour ω over not-ω. For example, suppose P hasn't given any thought to marriage, but if put in situations that gave her reason to consider it, the result would be a global utility for marriage (at the time of consideration) that is greater than the global utility for remaining unmarried. Pettigrew argues that vicarious transformative choice that pushes P in the direction of marriage isn't necessarily objectionable, for she would be glad that she was effectively manoeuvred into recognizing something that in a sense was already of value to her. In other words, it would be the right course of action as judged by herself. Note that this would be a case of what I referred to in §1.1.2 as a *weakly* transformative policy.

In the second case, P_{t1} has no global utility for ω, but if given the opportunity would set local utilities such that the sum would favour not-ω over

[45] Note that this is only a necessary condition for the permissibility of vicarious transformative choice on behalf of P_{t1}.

[46] Pettigrew 2023, §7.2.

ω. For example, if P were to encounter situations that cause her to consider the prospect of marriage for the first time, she would assign local utilities and weights such that the global utility for marriage is less than the global utility of remaining unmarried. Pettigrew argues that vicarious transformative choice that pushes P in the direction of marriage is impermissible in this case. His rationale is that

> If a society takes nudges of the sort we are considering to be legitimate, it permits governments to shape the ends of their citizens, and this, both liberal and libertarian agree, is beyond the pale.....granting a government such power is very likely to end in ways that are bad by the lights of the existing ends of the citizens[47]

In other words, the choice on behalf of P didn't bring her to recognize something that she in an important sense already valued ex-ante. It brought her to value something altogether new. Thus, even if she came to endorse being married ex-post, it was not the right choice as judged by herself ex-ante. This would an example of what I referred to in §1.1.2. as a *strongly* transformative policy. In Pettigrew's eyes, granting this power to the state will almost certainly end up running counter to what citizens presently care about.

The Global Utilities Approach is clearly designed for application within the political domain. However, like the Adaptive Preferences Test, it fails to come to grips with some distinctly political issues that bear on the permissibility of transformative policy. For the sake of argument, let us assume that Pettigrew's objection to strongly transformative policies is decisive - shaping the ends of its citizens is impermissible because it would allow the state to justify policies without reference to AJBT. As we've seen, weakly transformative policies can satisfy AJBT and so are not impermissible. However, it's not clear why, if we're not comfortable with the state shaping our normative commitments, we should be okay with its playing an active role in deciding how our normative commitments should be extended. People care about many things, so there are innumerable possibilities, all of which may satisfy Pettigrew's formulation of AJBT. If all of these possibilities are equally permissible as the aim of a weakly transformative policy, then this would seem to grant the state extraordinary latitude in determining the makeup of citizens' normative commitments. I am not claiming that this power shouldn't be granted to the state, only that its legitimacy is an unaddressed question at the level of political morality.

A second issue concerns weights. As we saw above, the global utility of a given outcome ω for persons P at time t1 is the sum of the weighted

[47] Pettigrew 2023, pp. 16–17.

local utilities that P_{t1} assigns (or would assign when prompted) to ω. In Pettigrew's analysis, the weights are assigned by P_{t1} in a way that reflects the degree of importance that he places on the local utility of ω for his past, present, and future selves in his deliberations. What's not clear from Pettigrew's analysis is why the state should always defer to these weightings for the purposes of determining whether a weakly transformative policy is permissible. Surely it's not uncommon for individuals to recognize that a possible course of action will, if undertaken now, be of great value to their future self, and yet discount this fact in deciding what to do. Many people do just this when deciding whether to quit smoking, for example. If there are societal impacts to people discounting the utility of their future selves in favour of their present self, then surely there is a case to be made for the state to assign different weights for the purposes of calculating the global utilities. Again, I make no claim either way here. The point is just that this is a distinctly political issue that is not captured by Pettigrew's framework.

A final issue is Pettigrew's objection to strongly transformative policy. As noted above, his worry is government overreach.[48] If we grant the state the power to fundamentally shape citizens' ends, the result will almost certainly be at odds with what they valued ex-ante. Now, one reply here is that this worry applies equally to the state possessing coercive powers. Since it is not a decisive objection in this case (except perhaps in the opinion of some anarchists), it's not clear why it's decisive when levelled against strongly transformative policy. There is a more fundamental issue, however. Just as we might question why the state should always take a person's weighting of local utilities as given, we might also question why it should always take a person's ex-ante assignment of local utilities as authoritative for the purposes of AJBT. Could there not be cases in which the state is permitted to to take the ex-post assignment of local utilities as authoritative for these purposes? Again, I make no claim either way here, but only point out that this is another distinctly political question that escapes Pettigrew's analysis.

1.2.3 Summary

My overall aim in this section has been to highlight a gap in current philosophical work on transformative experience when it comes to the political morality of vicarious transformative choice, i.e. transformative policy. To this end, I have discussed the nature of transformative experience, the puzzle of transformative choice and some approaches to its resolution, the moral

[48] Pettigrew 2023, p. 16.

salience of vicarious transformative choice, and two accounts of when such choices are morally permissible. While this has (I hope) demonstrated that there are many unanswered questions about the political morality of transformative policy, it is not clear that these questions have any practical significance. It is one thing to say that transformative policy means *could* be permissible. It is quite another to establish that there is any good reason for using them. If there are, as I argue in the next section, then we have reason to take transformative policy means seriously as tools of state action, rather than mere conceptual curiosities.

1.3 The Practical Case for Transformative Policy

Let us assume for the sake of argument that there are plausible conceptions of liberal political morality that can accommodate the use of transformative policy means. I would like to suggest that such means can have practical advantages that at least *prima facie* recommend their use.

1.3.1 Efficacy

The efficacy of policy depends on whether or to what degree it achieves its ends (impact) and how stable these results are over time (reliability). Impact is partly determined by the proportion of the target population that comes to regard themselves as having sufficient reason to act in accordance with the policy ends. Whether a given individual does so depends in part on their normative commitments.[49] Now, consider a policy that aims to raise voter turnout above 70% by levying a tax penalty against citizens who are eligible to vote but fail to do so. For the sake of simplicity, assume a population of four persons who differ only in the following ways:

> **Person 1**: Does not regard voting as a moral duty, has significant financial resources
>
> **Person 2**: Does not regard voting as a moral duty, has modest financial resources
>
> **Person 3**: Does not regard voting as a moral duty, has negligible financial resources
>
> **Person 4**: Regards voting as a moral duty

[49] Other relevant factors might include risk appetite, financial resources, education, etc.

The policy means are more likely to bring persons 2 and 4 to regard themselves as having sufficient reason to vote than persons 1 and 3, though for different reasons in each case. Person 2 is likely to vote because although they do not regard themselves as having a moral duty to do so, the burden that the tax penalty would impose given their modest financial resources gives them a prudential reason to do so. Person 4 is likely to vote because they regard themselves as having a moral duty to do so which is independent of the policy.[50] Person 1 is unlikely to vote because they do not regard themselves as having a moral reason, nor a prudential one in virtue of their significant financial resources. Person 3 is unlikely to vote because they do not regard themselves as having a moral reason to do so, nor a prudential one because they do not have sufficient financial resources for the tax penalty to be consequential for them (for example, they may not have enough income to file taxes, or they perform only low-paying cash jobs). The expected impact of the policy in this case is 50%, and so fails to achieve its ends.

Reliability is influenced by the intertemporal stability of the kinds of normative commitments that motivate persons to act in accordance with the policy aims. Different kinds of normative commitments exhibit different degrees of stability. For example, a person's values tend to exhibit greater intertemporal stability than their desires. This suggests that although persons 2 and 4 both cast votes at t_1, the likelihood that Person 2 does so at each of [t_2, t_3...t_n] is lower than for Person 4. For example, suppose that between t_1 and t_2, Person 2 secures a significantly higher paying job. This means that they effectively become Person 1. As a result, the expected impact of the policy drops from 50% at t_1 to 25% at t_2. This would not be the case for Person 4. Their regard for voting as a moral duty is not sensitive to income.

Suppose that instead of levying tax penalities against non-voters, the policy attempts to increase voter turnout through initiatives that are designed to bring people to regard voting as a moral duty for every citizen. If successful, these transformative policy means effectively change Persons 1-3 into Person 4. Those who didn't previously regard themselves as having a sufficient reason of any kind to vote would now have one (impact), and those whose reasons for voting were relatively unstable would now be motivated by reasons that are less sensitive to the contingencies of life (reliability). Though

[50] One could interpret the policy means as unsuccessful with respect to Person 4 because they do not act for the reasons encoded in the means. A more charitable interpretation is that the policy means do succeed insofar as they give Person 4 prudential reasons to vote that would be still be sufficient if they stopped regarding voting as a moral duty, and fails where this counterfactual fails. The former would be a case of Person 2, and the latter of Person 1 or 3. For the sake of simplicity I will assume the former case here.

omitting myriad complexities, this simple example highlights the potential advantages of transformative policy means in the face of impact- and/or reliability-related challenges that are rooted in the normative commitments of the target population.

1.3.2 Cost effectiveness

Certain kinds of policies are such that the efficacy of their means depends on the existence of monitoring and enforcement mechanisms, e.g. police services, criminal and civil courts, regulatory bodies, etc. These mechanisms are necessitated (in part) by the fact that individuals differ in both the content and weighting of their normative commitments. Consider two parents each of whom faces a choice of whether to send their child to school. Parent_1 regards themselves as having sufficient reason to do so, while Parent_2 does not (e.g. they think that it is better for the child to stay and work on the farm). In the absence of any policy, Parent_1 will send their child to school, while Parent_2 will not. Now, suppose the state implements a policy that requires parents to send their children to school or face criminal prosecution. For Parent_2 to regard themselves as having sufficient reason to comply, the sanction must not only be sufficiently weighty, but also be credible. It is compliance mechanisms that make the prescribed sanctions credible - for example, systems for monitoring registration and attendance of children and personnel, courts to adjudicate non-compliance disputes, etc. Notice that these systems are redundant when applied to Parent_1, who regards themselves as having sufficient reason to send their child to school that are independent of the reasons generated by the policy means. In an important sense, the costs associated with establishing and operating compliance systems are accrued solely in virtue of the content and/or weightings of Parent_2's normative commitments.

These considerations suggest a practical case for transformative policy means on grounds of cost effectiveness. Instead of emphasizing the use of coercive instruments to achieve the policy ends, suppose the policy means aim to influence Parent_2's normative commitments by offering temporary monetary incentives to motivate initial compliance paired with public outreach measures to reinforce the choice as the right thing to do for the child's future. If successful, Parent_2 would become effectively identical to Parent_1 - they would regard themselves as having sufficient reason to send their child to school that is independent of any requirement by the state that they do so. This would eliminate the need for, or at least significantly reduce the scale of, costly enforcement mechanisms. The practical justification for such policy means is therefore strongest when the costs of monitoring and enforcement are significant, the efficacy of these means are comparable to alternatives,

and it is undesirable or infeasible to abandon pursuit of the policy ends.

1.4 Why Transformative Manipulation?

Up to this point I have been concerned with showing that we lack a clear understanding of the political morality of transformative policy means, and that this gap is not merely of theoretical significance. However, my goal in this thesis is not to provide an a account of transformative policy means in general. Rather, I restrict my attention transformative manipulation specifically. Why this combination of mode and method? My primary motivation is that it presents us with a rare case of convergence amongst liberals on the *in-principle* impermissibility of certain kinds of policy means. In a sense, transformative manipulation is the hard case from a liberal perspective. If it can be shown to be permissible, then it suggests that other kinds of transformative policy means are as well.

On the one hand, it is undeniable that liberal states already utilize certain kinds of transformative policy means. The most obvious example is education policy. The very purpose of public education systems is to mould students to become productive members of society by helping them develop specific skills and internalize important social norms (if not values). However, neither the application of these policies nor their aims are hidden from students. Communicating the importance of attendance and effort for who they become in the future is part of their education. The method is therefore rational rather than manipulative. Indeed, if the state were to employ manipulative rather than rational methods in this context, the result would appear to be a system of indoctrination rather than education which, by its very nature, seems to be incompatible with any plausible conception of liberal political morality.

In a similar vein, prison rehabilitation programs aim to reduce recidivism through therapeutic services and vocational training. These programs are typically designed to not only help inmates develop useful skills, but also gain deeper insight into their own behaviour and character that (ideally) will help them become functional members of society upon their release. As in the previous example, these programs are conducted with the informed consent of the participants. That they understand the nature and purposes of their participation is a constitutive part of these programs. In contrast, "rehabilitation" (or "re-education") programs in authoritarian states often utilize techniques that are designed to induce uncertainty and confusion as a

means of affecting a change to the subject's beliefs and commitments.[51] That we don't typically observe this kind of 'rehabilitation' in liberal states reflects a tacit acceptance of transformative policy means within specific contexts, but only though rational engagement.

The same reticence about transformative manipulation can be observed in the other direction. Though still controversial, nudge policies - i.e. policies that are designed to help (but not force) people to make better decisions through the manipulation of choice situations[52] - have gained traction in recent decades. For example, in 2010 the UK government created the Behavioural Insights team, whose role is to leverage research on cognitive heuristics and biases to design nudge policies.[53] Three domains of application were initially proposed: Safer Communities (crime prevention, reducing anti-social behaviour, preventing 'degradation of surroundings'), The Good Society (pro-environmental behaviours, voter turnout, responsible parenting), and Healthy and Prosperous Lives (smoking, obesity, personal finances, education and training).[54] Crucially, however, the authors of the report emphasize the importance of public attitudes to the permissibility of such measures:

> Behavioural approaches embody a line of thinking that moves from the idea of an autonomous individual making rational decisions to a "situated" decision-maker, much of whose behaviour is automatic and influenced by their 'choice environment'. This raises the question: who decides on this 'choice environment'?...Policy-makers wishing to use these tools...need the approval of the public to do so.[55]

The legitimacy of manipulation at the level of choice scenarios is therefore a function of the degree to which it advances what members of the public

[51] Solzhenitsyn recounts the case of Erik Arvid Andersen, a member of a very wealthy Swedish family, who was kidnapped by Soviet agents in a failed attempt to turn him into a propagandist for the Soviet Union. Unable to persuade him, Soviet officials turned to other methods: "Since they didn't believe in his strength of mind, they locked him up in a dacha outside Moscow, fed him like a prince in a fairy tale... surrounded him with the works of Marx-Engels-Lenin-Stalin, and waited a year for him to be re-educated. To their surprise it didn't happen. At that point they quartered with him a former lieutenant general who had already served two years in Norilsk. They probably calculated that by relating the horrors of camp the lieutenant general would persuade Erik to surrender" (Solzhenitsyn 2018, p. 178).

[52] See Thaler and Sunstein 2008.

[53] Dolan et al. 2010.

[54] Ibid., p. 29.

[55] Ibid., p. 74.

already see themselves as having reason to do.[56] In other words, manipulative policy means are permissible only in the conservative mode. This limitation has been an important part of defences of nudge policies from the beginning. As noted in §1.2.2, Thaler and Sunstein's pioneering defence of nudge policies restricts their use to improving the lives of the target population as judged by themselves at the time of choice.[57]

Thus, just as the use of transformative means in liberal societies reflects an implicit prohibition on pairing them with manipulative methods, so too do those who defend manipulative means prohibit their use in the transformative mode. As mentioned above, this suggests a rare convergence amongst liberals on the limits of policy means, one that I argue is a mistake. As we will see, far from being prohibited by any plausible conception of liberal political, there is a plausible case to be made that the use of transformative manipulation as policy means is in fact *prescribed* by any plausible conception of liberal political morality under certain well-defined circumstances.

1.5 Conclusion

In this chapter I have sought to clarify the scope of my inquiry into the permissibility of transformative manipulation as a tool of public policy, and establish its philosophical and practical significance. I first provided a functional definition of policy means and differentiated between the constitutive features whose specification determines how this function is fulfilled. This served to highlight what makes policy means in the transformative mode distinctive. Next, I sought to demonstrate that although the philosophical literature on transformative experience and choice has begun to grapple with the morality of transformative choice for others, it has so far neglected the political morality of transformative policy means. I then argued that states have *prima facie* reasons for utilizing transformative policy means in the name of efficacy and/or efficiency under certain circumstances, thereby establishing the practical significance of such policy means. And finally, I clarified my choice to focus on transformative manipulation specifically, ar-

[56] From this condition is may seem to follow that nudge policies are not manipulative based on my definition of manipulation, i.e. intentionally obscuring either the application or the true aims of the policy means. This is not so. The subjects are not manipulated into endorsing the use of manipulative policy means; but the *application* of the means is manipulative.

[57] Pettigrew is an exception here. As we've seen in §1.2.2, he allows for weakly transformative nudges.

guing that it appears to be the most objectionable kind of transformative policy means from a liberal perspective.

In the next chapter, we consider precisely *why* transformative manipulation is objectionable from the perspective of any plausible conception of liberal political morality. This is the first step towards formulating the strongest possible argument against the permissibility of such policy means, an argument I seek to refute in later chapters.

Chapter 2

The Absolute Prohibition Thesis

One of the unifying characteristics of the liberal political tradition is a presumption that interference with individual freedom is wrong unless the interfering party can provide adequate justification for their actions. Call this the Fundamental Liberal Principle (FLP).[1] It is uncontroversial that certain kinds of policy means can satisfy this requirement. No one denies there are circumstances in which the use of persuasion or coercion (instrument), or appealing to prudential reasons (content), or rational engagement (method), or taking existing normative commitments as given (mode) can be justified. There are also policy means whose in-principle permissibility is a matter of debate. For example, there is no agreement amongst liberals about whether appeals to moral reasons (content), the use of manipulation (method), or modifying individuals' normative commitments (mode) can ever be justified. Transformative manipulation appears to be somewhat unique in this regard. It is an especially pernicious form interference, for its proximate target is not what persons do but who they are, and engages with them not as persons to be convinced but as things to be moulded. Such measures appear to be antithetical to the fundamental spirit of the liberal tradition. It is therefore hard to imagine any liberal theorist accepting that the use of transformative manipulation could ever be justified. Call this the Absolute Prohibition Thesis (APT).

The intuition that such policy means cannot be accommodated by any plausible account of liberal political morality is difficult to ignore, and so must be taken seriously. As a prelude to my positive argument for the per-

[1] Gaus 2005, p. 274.

missibility of transformative manipulation in later chapters, I would first like to identify the strongest possible argument for APT. The present chapter sets the stage for this task. In §2.1 I argue that a compelling case for APT must show both that transformative manipulation infringes on any plausible liberal conception of individual freedom, and that there are no sufficiently weighty reasons that could justify such restrictions. Further, there is an important connection between these requirements. Conceptions of freedom constitutively express evaluative judgements about the states of affairs they specify. These judgements play an important role in determining if and when interference with freedom so defined can be justified. In §2.2 I defend the claim that conceptions of freedom constitutively express evaluative judgements that establish the specified states of affairs as having a particular kind of normative salience. In §2.3 I argue in favour of a descriptive thesis about the nature of these evaluative judgements, namely, that they specify the conditions under which the value of agency is realized. The upshot is that a compelling case for APT must demonstrate that transformative manipulation interferes with something that all liberals agree is constitutive of conditions that realize the value of agency, and that the reason why it is constitutive of these conditions uniquely precludes any justification for the use of such policy means.

2.1 The Fundamental Liberal Principle

FLP is typically understood to be comprised of two claims: first, persons have no standing obligation to justify their actions. This is, in effect, a presumption of innocence. As Benn observes, "Unlike explanations, justifications...presume at least prima facie fault, a charge to be rebutted".[2] To say that no one has a standing obligation to justify their actions is therefore just to say that the default assumption is that their actions are permissible. And second, interference with others requires justification, and is otherwise morally wrong.[3] In other words, interference with individual freedom is at least *pro tanto* morally wrong. Therefore, the burden of justification for interference always falls on the interferer.

In the political context, FLP tells us that individuals are not responsible for demonstrating that they should not be subjected to state power; rather, it is up to the state to establish at least a prima facie case for the application of state power in a given case or range of cases. As a concrete example, consider the doctrine of probable cause as formulated in the Fourth Amendment of

[2] Stanley I. Benn 1990, p. 87.
[3] Gaus 2005, p. 274.

the United States Constitution:

> The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon **probable cause**, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized [emphasis added]

Prior to independence, officers of the Crown had statutory authority to detain persons and seize property based on simple suspicion of illegal activity.[4] However, first-personal attestation of mere suspicion is effectively incontestable by third parties. This evidentiary standard therefore places a de jure burden of justification on the state, but its extraordinary weakness means that the de facto burden falls on individuals to demonstrate to officers that they should *not* be subject to interference. The significance of probable cause as an evidentiary standard is that it must be demonstrated, and so is contestable by third parties. The Fourth Amendment therefore does much more than just strengthen the evidentiary standard. It transforms the state's de jure burden of justification into a de facto one as well, thereby establishing FLP as an effective restraint on state action.

If APT is correct, then transformative manipulation must fall within the scope of FLP. However, the proponent of APT must reckon with the fact that the liberal tradition encompasses myriad views on the nature of individual freedom, and therefore many different versions of FLP. To illustrate the significance of this challenge, it is helpful to draw on Gerald MacCallum's influential analysis of the concept of freedom. We will then be in a position to articulate just what a successful argument for APT must establish.

2.1.1 Freedom as a triadic relation

MacCallum argues that despite their myriad differences, conceptions of freedom are in fact different specifications of the same underlying concept.[5] The core of his argument is the claim that any meaningful statement about specific freedom[6] specifies a triadic relation between agents (x), "preventing conditions" (y), and "actions or conditions of character or circumstance" (z) of the form:

[4] Bodenhamer 2005.
[5] MacCallum 1967.
[6] Specific freedom is the freedom to ϕ, in contrast to overall freedom which is the aggregate of specific freedoms. See Carter 1999, pp. 12–14.

x is free from **y** to do or become **z**

The differences between conceptions boil down to different specifications of the range of possible values for one or more of each of the three variables.

Agents

The range of **x** determines the kinds of things that count as agents in the relevant sense, i.e. the kinds of things that are capable of moral and political freedom/unfreedom. According to what is undoubtedly the most influential account of agency, the 'Standard Model'[7], it consists of intentional action, understood as a complex of belief and desire: A's ϕ'ing is an intentional action when A desires some end, believes that ϕ'ing will help achieve that end, and on this basis executes on ϕ.[8] This account combines a conception of agency (intentional action) with a theory of action (belief-desire complex).[9] While the theory of action is now widely regarded as insufficient, there is general agreement that agency consists, at a minimum, of intentional action, and that the latter must be cashed out in terms of appropriately structured agent-specific states and events, which may include such things as beliefs, sources of motivation that constitute reasons (e.g. desires, values, etc), practical deliberation, practical commitments, and executing on these commitments. This analysis of agency lends itself to the view that only human beings are capable of genuinely intentional action. Velleman, for example, argues that genuinely intentional action ('full-blooded action') requires a capacity for higher order reflection on reasons - in other words, it is not merely goal-directed behaviour.[10] Insofar as this capacity is uniquely human, then only human beings count as agents.

But one needn't accept such restrictions on genuinely intentional action. Weaker notions that drop the need for higher-order reflection are perfectly coherent. Doing so greatly expands the kinds of things that qualify as agents in a freedom-relevant sense (e.g. certain non-human animals). Another way to expand the circle of agents is through accounts of non-reductive group agency.[11] These views regard certain groups as ontologically dependent on their members while also possessing a distinct kind of agency that is not reducible to that of their members. If one accepts some such account, then at least some groups are agents *qua* groups, and so can be free and unfree

[7] Velleman 2000.
[8] This view was independently developed by Anscombe 1957 and Davidson 1980.
[9] Schlosser 2019.
[10] Velleman 2000, p. 189.
[11] See List and Pettit 2011 and Tuomela 2013 for recent defences of such views.

in their own right. An even more radical way to break with the dominant conception of agency is to reject any necessary connection to intentionality. For example, Barandiaran et al. argue for a conception of agency that satisfies three criteria: there is a physical distinction between a system and its environment; the system must be capable of modulating the way it couples to its environment; and the system must have a baseline condition that modulation occurs with reference to.[12] Of course, such a move would produce an explosion of the kinds of things that count as agents (indeed, any living thing according to their account). Of course, it is doubtful that such a view meshes in a coherent way with other issues closely associated with agency, e.g. responsibility, dessert, etc.

I do not claim that the various ways of specifying the range of **x** are all equally plausible. Indeed, it is difficult to escape the intuition that the capacity for moral and political freedom is intimately bound up with richer notions of intentionality that involve what appear to be uniquely human capacities. What is important here is that choices about the kind of intentionality that qualifies something as an full-fledged agent determines those things for whom interference can be normatively salient *qua* interference. Some of these choices are clearly more plausible than others, but the point is that the range of **x** cannot be determined without making some such choices.

Preventing conditions

The range of **y** specifies what sorts of impediments count as restrictions on freedom. The scope of disagreement on this topic is enormous, so I restrict myself to describing only a few of its dimensions. Any plausible conception of freedom regards the actions of other persons as potential restrictions on freedom. Less obvious is whether natural impediments should count. Certainly, it would be bizarre to assert that there is no difference between an agent losing the ability to ϕ in virtue of a storm or volcanic eruption and losing the ability to ϕ due to the actions of other persons. But neither is it incoherent. Furthermore, there are other cases in which the line between agential and natural impediments is blurred - for example, where the effects of the latter could be prevented or remedied through human action.[13] It is exceedingly difficult to draw a sharp distinction between persons and natural phenomena as sources of unfreedom, for as Matthew Kramer observes,

> To come up with justificatory arguments that are not ultimately cir-

[12] Barandiaran, Paolo, and Rohde 2009.
[13] Gray 1990, pp. 22–23.

> cular, someone would have to adduce a more profound dichotomy that pertinently comprehends the distinction between the human and the natural. No obvious candidates for that role come to mind[14]

Even if we assume that only agent-based impediments can restrict freedom, there are still many other sources of disagreement about the range of $\mathbf{y}$. Theorists differ on whether, in addition to physical interference, promissory phenomena such as threats restrict freedom as well. Hobbes famously denies that they do, as do certain contemporary theorists such as Steiner and Nozick.[15] Joseph Raz, on the other hand, argues that threats can restrict freedom because, given certain assumptions about the nature of coercion, they can meaningfully limit an agent's available options.[16] Ian Carter, meanwhile, argues that while threats do not restrict specific freedom (i.e. the freedom to do this or that), they can reduce an agent's *overall* freedom by restricting the sets of compossible actions that are available to them.[17] For example, suppose my employer threatens to fire me if I miss work to attend a close friends funeral. This threat cannot *prevent* me from attending the service. However, it does restrict my access to any set of compossible actions that includes my attending the funeral and, say, being promoted from my current position at the company.

There is also the question of whether an impediment must render an action impossible, or merely make it more difficult. In other words, is the freedom to ϕ categorical or scalar? If the former, then we cannot meaningfully ask '*How* free is A to ϕ?', only '*Is* A free to ϕ?'. Oppenheim endorses the scalar view, as evidence by his claim that if we know that 70% of parking violations in a city are detected and punished, then we can say that '...drivers in that city are officially unfree to a degree of 0.7 to overpark and their freedom to do so is 0.3'.[18] Steiner on the other hand, argues that any statement about degrees of freedom is "...an elliptical abbreviation of a probabilistic judgement".[19] Suppose another city has an 80% detection and enforcement rate for parking violations. People in this city are not *less free* to overpark. Rather, says Steiner, they are are *less likely to successfully exercise* this freedom. Ian Carter and Matthew Kramer attempt a third way that captures

[14] Kramer 2003, p. 362.
[15] Hobbes 2003; Nozick 1974; Steiner 1994.
[16] Raz 1986, pp. 150–152.
[17] A set of actions are compossible if there is a possible world in which they all occur (Carter 1999, p. 180).
[18] Oppenheim 1961, p. 187.
[19] Steiner 1983, pp. 78–79.

intuitions on both sides of the question.[20] On the one hand, specific freedom - i.e. the freedom to do a specific thing - is categorical. Impediments that make it more difficult, but do not prevent, A from ϕ'ing do not therefore make A less free to ϕ. However, they do reduce A's *overall* freedom by diminishing A's set of compossible actions. What it means for an impediment to make ϕ'ing more difficult is that it in some sense makes it more costly for A to ϕ (whether in effort, time, money, etc). Thus, in order to ϕ, A will have to forgo doing other things that they would otherwise be able to do if doing ϕ were less costly. As a simple example, covering an oval racetrack in a thick layer of sand does not make a long distance runner less free to complete their five kilometre race. However, the additional effort required to run in sand might be such that they are unable to race again the next day due to exhaustion. The effect of adding the sand is to remove from their sets of compossible actions any set that includes both 'complete day one race' and 'complete day two race", thereby diminishing their overall freedom.[21]

Another dimension of disagreement about the range of $\mathbf{y}$ is whether exposure to the non-specific possibility of interference restricts freedom. The Republican political tradition answers in the affirmative when it comes to relationships of domination.[22] A has dominating power over B when A has the capacity to interfere on an arbitrary basis with certain choices that B is in a position to make.[23] In other words, A has unchecked power to interfere with all or some of B's choices as A sees fit. To use a canonical example, the fact that a slave-master chooses not to exercise his absolute power over his slaves does not seem to make them any less unfree. Perhaps they are better off than slaves who suffer a vicious master, but this is a fragile state of affairs. Their master could at any time and without any restraint inflict the most terrible abuses upon them. The Republican intuition is that these slaves are as unfree as those under a vicious master for the simple fact both groups are wholly dependent on the will of another.[24] Critics tend not to deny the relevance of these relationships to freedom, but question the analysis of domination as a distinctive kind of interference. Carter and Kramer again draw

[20] Carter 1999, pp. 232–233; Kramer 2003, pp. 169–174.

[21] See Miller 1984 for a critique of the compossible actions approach.

[22] Pettit 2002, Ch.2; Skinner 1998, Ch.2.

[23] Pettit 2002, p. 52.

[24] Quentin Skinner remarks that if "...you live under any form of government that allows for the exercise of prerogative or discretionary powers outside the law, you will already be living as a slave. . . . The very fact . . . that your rulers possess such arbitrary powers means that the continued enjoyment of your civil liberty remains at all times dependent on their goodwill. But this is to say that you remain subject or liable to having your rights of action curtailed or withdrawn at any time" (Skinner 1998, p. 70).

on their compossible actions framework to explain.[25] If domination restricts freedom, then there must be a non-trivial probability that the slave-master will exercise his power over the slaves.[26] Domination therefore reduces the probability that certain sets of compossible actions will be accessible to the dominated agent. In other words, the overall freedom of the slave is diminished to a degree that reflects the probability of the slave-master exercising his power. But, so the argument goes, this is just freedom as non-interference again.[27]

Up to this point we have been considering disagreement about external restrictions on freedom. There are myriad conceptions of freedom that also include certain kinds of cognitive or psychological phenomena within the range of $\mathbf{y}$. It is a defining feature of many kinds of mental illness that agents regard the symptoms as sources of interference in their lives. The obsessive-compulsive experiences the drive to enact certain behavioural rituals as beyond their control.[28] Phobias are typically recognized as irrational by those suffering them, but nevertheless cause them to avoid exposure to their objects. One might also regard false beliefs as potentially restricting freedom. Suppose a prison guard places Alice in a cell against her will but doesn't actually lock the door. However, having observed the guard going through the apparent motions of engaging the lock, Alice has no reason to believe that she could leave the cell at any time. The only thing seemingly preventing her from leaving the cell is therefore her false belief about being unable to leave.[29] Alternatively, one might argue that her belief doesn't make her unfree to leave - it merely leads her to fail to exercise what she is free to do.[30] Other candidates for internal constraints might include character traits, (e.g. cowardice, shyness, etc) or intellectual capabilities.[31]

[25] Carter 2008; Kramer 2008.

[26] On this point Carter remarks that "[I]t would again be a very unrealistic theory of politics that conceived of opportunities for the exercise of power as being accompanied, except in rare cases, by a trivially low probability of that exercise taking place (Carter 2008, p. 70).

[27] See Pettit 2008 and Skinner 2008 for replies to these and other worries about domination as a distinctive restriction on freedom.

[28] In the words of T.H. Green, such persons find themselves "...in the condition of a bondsman who is carrying out the will of another, not his own" 2011, p. 308.

[29] See Buchanan 2018 for a discussion of different ways of thinking about epistemic conditions of freedom.

[30] Kramer 2003, p. 266.

[31] For a general discussion of these options see Kramer ibid., pp. 264–271.

Ends

Finally, the range of **z** determines the kinds of things that one can be free or unfree with respect to. As above, the scope of disagreement on this topic is also enormous, so I will only touch on a few of its dimensions. A perennial source of disagreement between theorists is whether agents are free or unfree only with respect to actions, or, in addition, to the formation of one's character or self. Hobbes comes down in favour of the former. An agent is only free or unfree with respect to "...those things, which by his strength and wit he is able to do...what he has a will to".[32] Rousseau, in contrast, clearly defended the latter view, as evidenced by his idea of moral freedom as ... obedience to a self-prescribed law".[33] Both of these approaches continue to be influential in contemporary debate about the nature of individual freedom.

It is also possible to apply epistemic criteria to specify the range of **z**. For example, one might regard persons as free or unfree only with respect to rational ends, perhaps on Kantian-inspired grounds that connect one's status as a human being with reason, and reason with freedom. On this sort of view, I am neither free nor unfree with respect to ends that are incompatible with my nature as a rational creature because they aren't the kinds of things that a rational creature acting in accordance with its nature would choose. In other words, they are ends that are chosen only by stepping outside my nature as a rational creature in some sense, this nature being the very thing that renders me capable of freedom in the first place. Or one might argue on Hegelian-inspired grounds that connect freedom to living in accordance with a system of rational laws and institutions.[34] Those who reject this kind of analysis typically do so because it implies a concern not with the freedom of agents as we encounter them, but with, in Berlin's words, "...the free choice of [their] 'true', albeit often submerged and inarticulate, self".[35]

A somewhat connected debate concerns the application of moral or prudential criteria to determining the range of **z**. We see this in much ancient Greek thought that connects freedom (*eleutheria*) to a thing's essential nature (*phusis*). For both Plato and Aristotle, freedom

> ...refers to unrestrained flourishing "in accordance" with *phusis*, such that *phusis* is in fact what regulates development and gives it its own law, and where flourishing consists precisely in this agreement or con-

[32] Hobbes 2003, p. 146.
[33] Rousseau 1993, p. 167.
[34] Neuhouser 2000, Ch.3-5.
[35] Berlin 2002b, p. 180.

formity of the individual with the law of his essence[36]

Modern versions of this approach tend to draw on prudential reasons to restrict the range of **z**. Locke, for example, famously asserts "...that ill deserves the Name of Confinement which hedges us in only from Bogs and Precipices".[37] Though he later repudiated the view, Berlin argued in the initial version of *Two Concepts of Liberty* that freedom is the absence of obstacles to the fulfilment of one's desires.[38] The implication being, of course, that one is neither free nor unfree with respect to ends one does not desire. The modern liberal tradition has tended to eschew moral or prudential considerations as relevant to the range of **z**, however, instead arguing that agents are free or unfree with respect to whatever they do or could choose to do.

2.1.2 Two criteria for APT

As the previous section illustrates, the triadic relation is not a concept of liberal freedom specifically. There are limits to what a plausible conception of liberal freedom can accommodate in the range of each variable. For example, the range of **x** cannot be restricted to persons of a certain ethnicity or faith; the range of **y** cannot omit coercion; the range of **z** cannot include only morally permissible acts, etc. Nevertheless, this still leaves liberals with a great deal of room for interpretation. Consider the contrast between Mill and Kant on the range of **y**. According to Mill, individual freedom is diminished by "...compulsion and control, whether the means used be physical force in the form of legal penalties, or the moral coercion of public opinion".[39] In other words, the range of **y** is made up of external influences that impede or otherwise frustrate our ability to pursue our desires. For Kant, this is only part of the story, for freedom also concerns determinates of the will itself. A person with a heteronomous will - that is, a will that is determined by influences outside of itself - is no more free than one whose actions are restricted by others. However, since everything in nature is determined by causal laws, freedom cannot require an undetermined will. Rather, it must be self-determining or autonomous.[40] It follows that for Kant, the range of **y** contains not only external sources of interference, but also certain kinds

[36] Romano 2004, p. 252.
[37] Locke 2013, p. 305.
[38] Berlin 2002a, pp. 30–31.
[39] Mill 2003, p. 80.
[40] For Kant, this of course means the self-imposition of a moral law. See Ware 2023, §2.1 for a general discussion of the role of heteronomy and autonomy in Kant's account of freedom.

of internal influences such as compulsions or overly strong desires that are capable of rendering the will heteronomous.

Recall that FLP rules out the use of transformative manipulation in principle only if its use necessarily undermines individual freedom. In light of widespread disagreement amongst liberals about the nature of freedom - particularly the ranges of **y** and **z** - a compelling argument for APT must therefore demonstrate that

1. Transformative manipulation interferes with individual freedom on any plausibly liberal specification of the triadic relation

2. There are no sufficiently weighty reasons that could overcome the liberal presumption against this kind of interference

As it turns out, (1) and (2) are more closely connected than they may at first appear. Specifications of the triadic relation are not arbitrary. Any choice of what is and is not included in the range of each variable is a product of evaluative judgements about what sorts of agents, preventing conditions, and ends *matter* in some sense. Now, the nature of these evaluative judgements plays an important role in determining the limits of justification for interferences with freedom so defined. Or, put another way, conceptions of freedom express evaluative judgments that affect the kinds of considerations that can, and more importantly cannot, justify infringements upon it. As we'll see in Chapter 3, this fact plays a crucial role in explaining why, unlike the other kinds of policy means covered in Chapter 1, transformative manipulation seems to be uniquely incompatible with the liberal tradition.

The claim that conceptions of freedom express evaluative judgements is not an uncontroversial one, however. Certainly, no one denies that we make evaluative judgments about the value *of* freedom (once defined), but this is different from saying that the definitions themselves express certain kinds of evaluative judgements. In the next section I argue that conceptions of freedom are constitutively evaluative, before moving on to elucidate the content of the relevant judgments.

2.2 The Value-Dependence of Freedom

There is a clear sense in which we speak of agents and objects as being free to do this or that (or to this or that degree) without making evaluative judgments of any kind. In ordinary usage such claims may be about capabilities, e.g. to say that a tree is not free to pull itself out of the ground and stroll around is to say that it lacks the ability to do so under any circumstance.

More commonly, these claims describe physical impediments to what something or someone would otherwise do or be able to do, e.g. by virtue of its banks, the river is not free to spread over the land (to take an example from Hobbes)[41]; or, the prisoner is not free to leave their cell because the door is locked. It is this sort of usage that Hillel Steiner has in mind when he observes that

> When we ask whether a person is free to do a particular action, we typically don't imagine ourselves to be asking an evaluative question. Rather, we're asking a factual question, the (affirmative) answer to which is presupposed by any evaluative question about his doing that action[42]

To say that freedom is x in this sense is to provide a stipulative definition of the term, i.e. attach the label 'freedom' by fiat to the states of affairs described by x. Conceptions of moral and political freedom are not like this, however. Philosophers routinely disagree about the nature freedom in this sense, even going so far as to suggest that their interlocutors fail to articulate a genuine conception at all. Consider Berlin's caustic allusion to Rousseau in his remark that

> Hobbes... did not pretend that a sovereign does not enslave; he justified this slavery, but at least did not have the effrontery to call it freedom[43]

This sort of claim would be at best deeply confused and at worst meaningless if conceptions of moral and political freedom were just stipulative definitions. To claim that a stipulative definition has gotten something wrong is to misunderstand the nature of stipulative definitions. The thesis I would like to defend here is that conceptions of moral and political freedom express evaluative judgements about the conditions - as defined by a specification of the triadic relation - under which a certain kind of value is realized, and so are constitutively normative. For example, to say that freedom is non-interference is to say that a certain kind of value is realized when there are no external constraints on an agent's ability to pursue their aims; to say that freedom is non-domination is to say that a certain kind of value is realized when persons are not subject to arbitrary power; to say that freedom is self-mastery is to say that a certain kind of value is realized when persons play

[41] Hobbes 2003, p. 145.
[42] Steiner 1994, p. 11.
[43] Berlin 2002a, p. 210.

an active role in shaping their own character; and so on. More generally, to say that freedom is x is to say that the states of affairs described by x have a particular kind of normative salience that warrants the appellation 'freedom'.[44] Call this the *value-dependency thesis*.[45]

It is important to distinguish the claim that conceptions of freedom are value-dependent from the claim that they are *value-laden*.[46] A value-laden conception is one that makes use of explicitly evaluative terms in the definition. Consider the following examples:

> ...the appropriate condition for regarding an obstacle as a constraint on freedom is that some other person or persons can be held morally responsible for its existence.[47]

> ...freedom in all the forms of doing what one will with one's own, is valuable only as a means to an end. That end is what I call freedom in the positive sense: in other words, the liberation of the powers of all men equally for contributions to a common good[48]

The appeal to moral responsibility in the first passage limits the kinds of interferences that undermine freedom - that is, the range of **y**. In effect, it is not the absence of external interference that is necessary for freedom, only the absence of *unjust* interference. In contrast, the appeal to a common good in the second passage restricts the kinds of aims that persons can be free or unfree to pursue or achieve - that is, the range of **z**.

The value-dependency thesis does not entail that conceptions of freedom are value-laden (nor vice versa). I highlight this point for two reasons. First, value-laden conceptions of freedom are subject to a number of powerful critiques. Most obviously, they produce judgements that are radically at odds with our basic intuitions about the nature of freedom and unfreedom. For example, suppose that agents are only free or unfree with respect to virtuous aims. It follows that locking up someone who would otherwise engage in

[44] As William Connelly observes, "In the ordinary language of political life and in more formal systems of political inquiry the normative dimensions in the idea of freedom are not attached to it as "connotations" that can be eliminated; without the normative point of view from which the concept is formed we would have no basis for deciding what "descriptive terms" to include or exclude in the definition" (Connolly 1993, p. 141). See also Benn and Weinstein 1971, p. 195.

[45] I draw this terminology from Carter 2015, p. 285. See also Kramer 2018, p. 376.

[46] Carter 2015, p. 284.

[47] Miller 1984, p. 190.

[48] Green 2006, p. 372.

nothing but vice does not make them unfree to do anything.[49] Conflating value-dependence with value-laden-ness can easily lead one to mistakenly apply these critiques to the value-dependency thesis. Second, even a cursory glance at the liberal tradition reveals myriad value-free (i.e. not value-laden) conceptions.[50] If value-dependence entailed value-laden-ness, then the value-dependency thesis could be rejected on purely factual grounds.

There are at least two ways one might attempt to refute the value-dependency thesis, however. The first, which can be quickly dismissed, is to arbitrarily specify the range of each of the three variables in the triadic relation. This would not produce a conception of moral and political freedom, however. If one cannot provide any justification for the content of one's conception, then it's not clear how it could carry any weight in our moral or political deliberations. I therefore leave this strategy aside.

The second strategy, is to argue that we can provide non-arbitrary specifications of the triadic relation without recourse to evaluative judgements about the conditions under which some kind of value is realized. Several prominent theorists purport to do so by offering an explicative definition of freedom - that is, one that sharpens certain existing patters of usage by applying purely methodological or theoretical criteria (e.g. parsimony, clarity, etc).[51] Such efforts do make implicit value judgments about the virtues of certain methodological criteria, but these are not judgements about the conditions under which other more substantive kinds of value are realized.

As an illustrative example, let us look at Felix Oppenheim's account. He argues that an adequate definition of (social) freedom must "... explicate the concept it defines, must be operational [i.e. empirical/measurable], must be fruitful, and must be valuationally neutral".[52] By employing these methodological criteria, he formulates a conception according to which an agent A is free to ϕ iff no other agent (a) prevents A from ϕ'ing; (b) makes it necessary for A to ϕ; or (c) makes it punishable for A to ϕ or to not-ϕ.[53] This definition, he argues, clarifies "... what is generally entailed by such vague terms as 'liberty' or 'free', as they occur in everyday language, and more particularly in political writings".[54] Thus, we have a non-arbitrary specification of the triadic relation that does not rest on evaluative judgements about the

[49] For a more detailed discussion of this and other critiques, see Kramer 2003, pp. 101–103.
[50] Mill, for example, does not make use of explicitly evaluative terms in his definition of freedom.
[51] See Kramer 2003, pp. 152–153 for a discussion of these efforts.
[52] Oppenheim 1985, p. 6.
[53] Ibid., p. 6.
[54] Ibid., p. 6.

conditions under which any kind of substantive value is realized

Oppenheim does not directly appeal to any substantive evaluative claims to justify his definition; however, his definition is value-independent only if the usage that he seeks to explicate is itself value-independent. There is good reason to deny that it is. If ordinary usage of the expression in moral and political contexts did not reflect substantive evaluative judgements about the conditions under which some kind of value is realized, then it would be a mystery as to why the relevant conditions would be regarded as having the kind of normative salience that speakers so clearly regard them as having. In other words, ordinary speakers do not typically justify the conceptual boundaries of freedom as they understand it by exclusive appeal to methodological or theoretical considerations. There are usually much more substantive reasons for their allegiance to a particular conception of freedom. Oppenheim's application of purely methodological criteria may sharpen these value-dependent usages of the term – say, by purging them of aspects that are not subject to empirical measurement – but doing so does not strip them of their value-dependence. His explicative definition of freedom is therefore not value-independent. At best, it is a value-dependent product of a value-independent procedure for adjudicating between or sharpening different patterns of ordinary usage.

The value-dependent nature of explicative definitions of freedom is also apparent in Matthew Kramer's work. He is at pains to divorce his account of freedom from substantive evaluative claims, remarking that

> My efforts to explicate the concept of freedom are not driven or shaped by a political vision that might be distinctively served by those efforts. Instead, the objective herein is to elucidate and hone an array of concepts that will enable greater rigour in the discussion of political affairs, whatever one's stances in respect of those affairs might be[55]

Like Oppenheim, Kramer starts from certain patterns of ordinary usage and applies methodological and theoretical desiderata to arrive at his explicative definitions of freedom and unfreedom:[56]

1. An agent is free to ϕ iff they are able to ϕ

2. An agent is unfree to ϕ iff the following conditions are met

   (a) They are directly or indirectly prevented from ϕ'ing by another person

[55] Kramer 2003, p. 152.
[56] Ibid., p. 3.

(b) They would be able to ϕ if (a) did not obtain

This definition is value-dependent in virtue of (2a), which restricts the range of **y** to direct or indirect interference by persons. According to Kramer, natural phenomena that no one could reasonably anticipate or have influence over are excluded because "When our focus is on socio-political freedom...we have to give some recognition to the distinctive importance of human interrelationships as sources of restrictions on that freedom".[57] This is a perfectly reasonable claim, one that is hard to imagine anyone disputing. However, this does not obviate its status as an evaluative judgement that is neither methodological nor theoretical. Like Oppenheim, Kramer's explicative definition of freedom only appears to be value-independent because the substantive evaluative judgements are embedded in the object of analysis rather than within the analysis itself.

Earlier I remarked that conceptions of freedom express substantive evaluative judgements that affect the kinds of considerations that can justify infringements upon it. My aim in this section has been to defend the first part of this claim by arguing that conceptions of moral and political freedom are constitutively value-dependent. In order to defend the second part of the claim, it is necessary to clarify the nature of these evaluative claims, and indeed whether we can make any general claims about them at all. I take up these issues in the following section.

2.3 Freedom and Value of Agency

If the value-dependency thesis is correct, then conceptions of freedom are not merely specifications of the triadic relation. Rather, any specification of the triadic relation is an expression of certain evaluative judgements that justify the chosen range of possible values for each of the three variables. To say that freedom is x is therefore to say that some kind of value is realized when the specified states of affairs obtain. But what kind(s) of value? In this section I argue that conceptions of freedom specify conditions that realize the value of agency. In particular, they specify (i) what kinds of things count as agents of the relevant kind; (ii) the nature of the ends in virtue of which agency is valuable; and (iii) the kind of relation between agents and the relevant ends that, when it obtains, realizes this value.

This thesis has two features that require some explanation. First, it is a purely descriptive thesis. I do not claim that this is how we *should*

[57] Kramer 2003, p. 362.

think about moral and political freedom. As we'll see, this understanding is already implicit in existing accounts. My reason for taking the descriptive route is that APT is meant to establish that the liberal tradition already rules out the use of transformative manipulation. It is not my aim to prescribe fundamental revisions to the foundations of liberalism in order to rule out the use of such policy means, only to consider whether it already does so. However, one might wonder if my thesis can be rejected right out of the gate in light of this clarification. It is arguably a core feature of the liberal tradition that it regards agency as having non-derivative value of some kind. That is to say, liberals do not typically regard agency as having merely instrumental value. But doesn't (ii) - i.e. that conceptions of freedom specify aims in virtue of which agency is valuable - entail that the value of agency is wholly derivative of the value of the specified ends? If so, then my claim that conceptions of freedom specify conditions that realize the value of agency cannot be correct as a descriptive thesis. This is not so. From the fact that A is valuable *in virtue* of B does not entail that the value of A is *derivative* of the value of B. Consider the following examples:

(a) Michelangelo's frescoes in the Sistine Chapel are only valuable because they draw tourists whose money contributes to the upkeep of the Vatican Museums

(b) Michelangelo's frescoes in the Sistine Chapel are valuable because they are objects of worthwhile aesthetic experience

In (a), the value of the frescoes is wholly derivative of the value of the Vatican Museums more generally. This is the instrumental reading of 'in virtue'. In (b) the frescoes have non-derivative value as sources of aesthetic experience. In other words, the frescos are not a *means* to having a valuable aesthetic experience - to apprehend them is to have a valuable aesthetic experience. In this sense, they are intrinsically valuable. Notice, however, they would not be valuable in this way if the world contained nothing that was capable of having aesthetic experiences. Their having value depends on the existence of things that are capable of engaging with them in the appropriate way, but the value of the frescos is not derivative of the value of these things.[58] To say that conceptions of freedom specify aims in virtue of which agency is valuable is therefore to say either that agency is valuable merely as a means to realizing other kinds of value connected to these aims, or that it is valuable simply because the universe contains worthwhile aims. This latter view is

[58] See Raz 2004, pp. 148–151.

perfectly consistent with views that regard agency as having non-derivative value.

The second point of clarification is that my thesis is about extant conceptions of moral and political freedom generally, not just liberal conceptions. This may seem unnecessarily strong given that APT concerns the liberal tradition specifically. However, the stronger thesis has the virtue of helping to explain not only why liberal conceptions of freedom seem to rule out the use of transformative manipulation in principle, but also why liberal conceptions of freedom seem to be unique in this regard when compared with non-liberal conceptions. So while the weaker thesis is sufficient for my purposes, the stronger one is more illuminating.

Now, the claim that conceptions of freedom specify conditions that realize the value of agency is plausible as a descriptive thesis only insofar as it is borne out by extant conceptions of freedom. While a wide-ranging survey would be ideal, this would require far too much space to accomplish here. Instead, I focus on two examples, namely, Hobbes and Rousseau. Part of the motivation for this selection is that they conceive of freedom in profoundly different ways. If my thesis holds true of their accounts, then this is good evidence that it holds more generally. The selection is also motivated by the fact that although neither Hobbes nor Rousseau comfortably fall within the liberal tradition, each of them is arguably the progenitor of a distinct strand of liberal thought about the nature of freedom. Hobbes' concern with external impediments to action is found in the writings of Locke, Bentham, and Mill. Meanwhile, Rousseau's emphasis on relevance of internal constraints to freedom was picked up by the likes of Kant, Green, and Dewey. Using their views as test cases therefore has the advantage of illuminating the kinds of considerations that influenced numerous liberal theorists in their thinking about the nature of freedom.

I begin by examining Hobbes' conception of freedom before turning to Rousseau's. For the sake of brevity I omit discussion of (i) what counts as an agent of the relevant kind, and instead focus my attention on demonstrating that their accounts are motivated by claims about (ii) the kinds of ends in virtue of which agency is valuable, and (iii) the kind of relation between agents and the relevant aims that, when it obtains, realizes this value.

2.3.1 Hobbes: freedom and felicity

Hobbes provides a number of definitions of freedom. Arguably his clearest formulation is this: "A FREE-MAN, is he, that in those things, which by his strength and wit he is able to do, is not hindred to doe what he has a will

to".[59] On this view, there are two sets of aims that agents are neither free or unfree with respect to: those that are not within their power to pursue or achieve, e.g. I am neither free nor unfree to live forever, or (according to Hobbes) to determine my will, desires, or inclinations; and those that I do not will to pursue, e.g. I am made no more or less free by being incarcerated unless I will to do something that requires me to leave my cell.[60]

To see how this definition reflects claims about the aims in virtue of which agency has value, we must consider Hobbes' conception of the good.[61] On the one hand, what we customarily call good is that which we desire because desire is a source of pleasure, both in its satisfaction (pleasures of sense) and in the anticipation of doing so (pleasures of the mind).[62] However, what is truly good is that which contributes to a life containing the greatest possible amount of ongoing pleasure – that is, one that is maximally felicitous. Hobbes maintains that felicity requires both kinds of pleasure. The pleasure derived from the satisfaction of a desire is too transient to sustain ongoing pleasure, since it lasts only so long as its object engages with the senses. It is the anticipation of satisfying our desires – i.e. pleasures of the mind – that makes the greatest contribution to felicity. However, this anticipatory pleasure requires confidence in one's power to satisfy the desire (i.e. hope). Absent this, it is a source of frustration and pain. Pleasures of sense play a critical role here. Hope comes from an awareness of the power to satisfy one's desires, but this awareness develops only through having successfully done so in the past. In this way, pleasures of sense facilitate felicitous pleasures of the mind. Thus, if we accept that agency has value only if the universe contains worthwhile aims, and worthwhile aims are picked out by one's conception of the good, then on the Hobbesian view agency has value by virtue of the existence of felicitous pleasures of sense and of the mind.

That Hobbes' conception of freedom is an expression of his view of the good is borne out by the fact that it excludes all and only those aims that cannot, in principle, be felicitous. A desire that is not within one's power to satisfy cannot produce pleasures of sense. And while an agent who mistakenly believes that it is within their power to satisfy will experience pleasures of the mind in the short-run, its unsatisfiability becomes a source of pain in the long-run.[63] Similarly, things that an agent does not will to do produce neither

[59] Hobbes 2003, p. 146.
[60] Ibid., pp. 145–146.
[61] There are a variety of interpretations of Hobbes' position on these topics, but for illustrative purposes I rely on Abizadeh 2018, Ch.4.
[62] Hobbes 2003, p. 39; Abizadeh 2018, p. 146.
[63] Abizadeh 2018, pp. 160–161.

pleasures of sense nor of the mind. Since the will is "... the last Appetite [i.e. desire], or Aversion, immediately adhaering to the action, or to the omission thereof", a desire not willed is one that is not pursued, and so will neither be satisfied nor create any expectation of its satisfaction.[64] This is not to say that any and all aims that satisfy Hobbes' definition are in fact felicitous, only that they are (at least in principle) potentially so.

As mentioned above, conceptions of freedom also specify the relation between agents and worthwhile aims that, when it obtains, realizes the value of agency. In the Hobbesian picture, it is a relation of unpreventedness: an agent is free to do that which they have willed and is in their power to do if and only if they are not subject to external impediments to their doing so.[65] If the analysis up to this point is correct, then this is equivalent to the claim that the value of agency is realized when agents face no external impediments to pursuing potentially felicitous aims.

This characterization raises two questions. First, given that agency is valuable in virtue of the existence of actually felicitous aims, why is its value realized with respect to aims that are only potentially felicitous? The answer is found in Hobbes' observation that whether or not an aim is felicitous depends on the balance of pleasures and pains produced by its long-run consequences.[66] But few (if any) persons are capable of apprehending the full chain of consequences that flow from a given action. The best we can do is act on those aims that are in our best estimation felicitous – that is, on what is *apparently* good.[67] Agency is valuable in virtue of objective facts about the world, but the aims relative to which this value is realized reflect the epistemic limitations we face in determining these facts.[68]

Second, why is the relevant relation between agents and potentially felicitous aims one of non-interference rather than achievement or satisfaction? The simple answer is that the latter is inconsistent with his conception of the good. Suppose that A wants (and has the means) to buy a boat because she believes it will be a source of net pleasure for her in the long-run. If the value of agency is realized by *satisfying* potentially felicitous aims, then it is realized only when she buys the boat. But suppose her belief is mistaken

[64] Hobbes 2003, p. 44.
[65] Ibid., p. 91.
[66] Baumgold 2017, p. 50.
[67] Ibid., p. 50.
[68] On this point, Abizadeh maintains that "A gap therefore appears in Hobbes's ethics between his substantive theory of the good and his theory of reasons: the latter but not the former is relativized to the epistemically accessible evidence" Abizadeh 2018, p. 166.

– perhaps the operational and maintenance costs impose pains on her that outstrip the pleasure of boating. In this case, we would be forced to say that the value of agency is realized even though she does something that *detracts* from felicity. But this is hardly coherent given Hobbes' conception of the good. It would entail that although agency is valuable in virtue of the existence of felicitous aims, it is possible to realize its value by satisfying those that are infelicitous.

To sum up, Hobbes' conception of freedom is an expression of two general claims: first, agency is valuable in virtue of the existence of felicitous aims; and second, this value is realized when agent's face no external barriers in their pursuit of aims that are in their best estimation felicitous. We now turn to Rousseau's conception of freedom which, although very different in substance, is nevertheless likewise concerned with the conditions under which the value of agency is realized.

2.3.2 Rousseau: freedom and virtue

For Rousseau, an agent is free when they are subject to no will but their own.[69] However, there is more built into this idea than first appears; he draws a crucial distinction between independence of the will and freedom:

> We should not confuse independence with freedom. These two things are so different that they are even mutually exclusive. When each one does what pleases him, he often does what displeases others, and it would be wrong to call that a state of freedom. Freedom is not so much the realization of one's will as independence from the will of others, and it does not involve making another's will dependent on one's own. Anyone who is master cannot be free, and to reign is to obey[70]

Freedom does not consist merely in being subject to no will but one's own, but also in refraining from imposing one's will on others. For this reason, it is perhaps more accurate to describe Rousseau's conception of freedom as co-independence of the will. This conception appears in several guises in Rousseau's writings, most notably as natural, civil, and moral freedom. These are not different conceptions of freedom per se, but rather context-specific instantiations. As we'll see, their context-specificity is a function of Rousseau's view of human nature, the malleability of which influences the

[69] "Freedom does not consist so much in doing one's will as in not being subjected to the will of others" (as cited in Neuhouser 2000, p. 69).

[70] As cited in O'Hagan 1999, pp. 68–69.

aims in virtue of which agency is valuable, and the relation that realizes this value.

Natural freedom

In the pure state of nature, the freedom of persons consists of "... an unlimited right to anything that tempts him and that he can attain... limited only by the powers of the individual".[71] This formulation most fundamentally reflects Rousseau's view of human nature prior to the formation of social groups. Natural man is an unreflective and largely solitary creature who is motivated wholly by amour de soi, a self-love concerned only with securing one's own wellbeing, which for natural man consists merely in the satisfaction of his basic physical needs. Crucially, however, his pursuit of these aims is tempered by a sense of pity that produces "... an innate abhorrence to see beings suffer that resemble him".[72] Against a backdrop of material abundance found in the pure state of nature, these characteristics ensure that natural man has neither the need nor the desire to impose his will on others, nor they on him. His is a natural freedom because when he is left to his own devices, his motivations are such that they are consistent with co-independence of the will.

Based on the formulation above, the aims relative to which natural man is free or unfree are wholly determined by his physical needs. This is because, driven by amour de soi, he desires only that which is necessary to satisfy these needs, and their satisfaction requires nothing that he cannot obtain. For Rousseau, natural man's highest good is happiness or contentment.[73] As the following passage illustrates, his happiness is contingent only on the satisfaction of these aims:

> As long as men remained satisfied with their rustic huts; as long as they were content with clothes made of the skins of animals, sewn with thorns and fish bones; as long as they continued to consider feathers and shells as sufficient ornaments, and to paint their bodies different colors, to improve or ornament their bows and arrows, to fashion with

[71] Rousseau 2002b, p. 167.
[72] Rousseau 2002a, p. 106.
[73] The centrality of happiness to Rousseau's conception of the good comes out in myriad passages throughout his writings. A particularly clear example is found in *Emile*, wherein the Savoyard Priest asserts that "Supreme happiness consists in self-content; that we may gain this self-content we are placed upon this earth and endowed with freedom, we are tempted by our passions and restrained by our conscience" (Rousseau 1979, p. 281).

> sharp-edged stones some little fishing boats, or clumsy instruments of music; in a word, as long as they undertook such works only as a single person could finish, and stuck to such arts as did not require the joint endeavors of several hands, they lived free, healthy, honest and happy, as much as their nature would admit, and continued to enjoy with each other all the pleasures of an independent intercourse... [74]

Recall that agency is valuable because the world contains worthwhile aims, and worthwhile aims are picked out by one's conception of the good. From this, it follows that the aims relative to which natural man is free or unfree (i.e. that which he desires and is within his power to obtain) are identical to those in virtue of which his agency is valuable (i.e. those that are conducive to happiness).

If the forgoing analysis is sound, then Rousseau's conception of natural freedom asserts that the value of natural man's agency is realized when his ability to pursue aims that are conducive to his happiness is not constrained by foreign wills. However, as noted in the context of Hobbes' conception of freedom, there is something puzzling about the idea that agency is valuable by virtue of the existence of worthwhile aims, and yet realizing its value does not require their satisfaction. The resolution of this puzzle in Rousseau's case is characteristically idiosyncratic. It is a peculiar consequence of his description of natural man and the pure state of nature that being subject to no will but his own is in effect a sufficient condition for his satisfying worthwhile aims. In other words, when natural man's ability to pursue worthwhile aims is not constrained by foreign wills, he effectively never fails to satisfy them.

Moral and Civil Freedom

Exiting the pure state of nature marks a transformation of human nature itself, which in turn transforms the conditions under which freedom obtains. Persons outside of the pure state of nature ('civilized man') differ from natural man in two important respects. First, for civilized man, the purview of amour de soi extends beyond mere self-preservation. The emergence of social relationships catalyses a richer conception of wellbeing by transforming his self-understanding and giving rise to new wants. And second, these new circumstances give rise to amour-propre, a distinctly comparative form of self-love that seeks esteem or approval from others. Though amour-propre is neither inherently good or bad, when left unchecked – or 'inflamed' - it produces sentiments such as vanity, envy, and pride that erode pity and

[74] Rousseau 2002a, pp. 119–120.

"... [render] men avaricious, wicked, and ambitious".[75] Natural freedom is thereby dissolved in a sea of conflict that creates and exacerbates inequalities between persons, and in time motivates the establishment of unjust societies that codify these inequalities in the name of stability. The bulk of humanity is thereby condemned to a life of "... perpetual labor, servitude, and misery".[76]

Civilized man cannot regain natural freedom, for his nature is irrevocably changed.[77] For him, co-independence of the will requires dealing with the pernicious influence of inflamed amour-propre. The first dimension of this struggle concerns civil freedom, which obtains under the protection of civil laws from those who, driven by inflamed amour-propre, would seek to interfere with others' ability to pursue lawful aims.[78] This definition may seem to imply a contradiction: undermining independence of the will via law does not undermine freedom (co-independence of the will). That Rousseau was acutely aware of the problem is evident from the task he sets for himself in The Social Contract, namely, to

> ... find a form of association that may defend and protect with the whole force of the community the person and property of every associate, and by means of which each, joining together with all, may nevertheless obey only himself, and remain as free as before[79]

His solution is to reconcile freedom and law via the General Will.[80] Roughly speaking, freedom and law are compatible only if the latter is an expression of the will of each and every person that is subject to it, i.e. the General Will. As noted above, however, inflamed amour-propre motivates the pursuit of private interests, which brings persons into conflict with one another and so undermines freedom. The General Will cannot, therefore, be an aggregation of what persons actually will. Rather, it expresses what persons rationally will in the knowledge of what is conducive to their own good, namely, the common interest or virtue.[81] In being subject to the General Will persons are therefore subject to nothing more than the truest expression of their own will. It is this fact that gives meaning to Rousseau's infamous assertion that

[75] Rousseau 2002a, p. 123.
[76] Ibid., p. 125.
[77] Ibid., p. 144.
[78] Rousseau 2002b, p. 167.
[79] Ibid., p. 163.
[80] Ibid., p. 175.
[81] We see this in Rousseau's remark that "By themselves, the people always desire what is good, but do not always discern it" (ibid., p. 167).

"... whoever refuses to obey the general will shall be constrained to do so by the whole body; which means nothing else than that he shall be forced to be free".[82]

The second dimension of the struggle for freedom concerns a distinctly internal threat posed by inflamed amour-propre. As noted above, the unchecked desire for recognition or esteem very easily generates sentiments such as envy, resentment, and pride that are corrosive to amour de soi, and in particular, our sense of pity. This conflict between socially acquired vice and natural goodness not only creates conflict between persons, but also divides persons against themselves.

> Swept along in contrary routes by nature [amour de soi] and by men [amour-propre], forced to divide ourselves between these different impulses, we follow composite impulse which leads us to neither one goal nor the other. Thus, in conflict and floating during the whole course of our life, we end it without having been able to put ourselves in harmony with ourselves and without having been good either for ourselves or for others.[83]

A divided will undermines freedom in two ways: first, the dictates of inflamed amour-propre run counter to those of amour de soi, which means that exercising the will is very often an act of self-subjugation.[84] And second, inflamed amour-propre takes one's status relative to others as its object and thereby yokes one's passions to their will, undermining one's own independence. The solution to this problem is moral freedom, defined as "... obedience to a self-prescribed law". That is to say, obedience to a self-prescribed moral law that brings amour-propre into harmony with amour de soi. There is only one such law: align one's will with the General Will, and therefore, virtue.[85] Virtuous persons channel amour-propre and reason to extend their natural compassion for individuals to humanity as a whole. In doing so, they recognize that the happiness of each depends on the happiness of all, and so will only that which is consistent with the common interest.[86] The internal threat to independence of the will is thereby defused when amour-propre is bent towards virtue by the hand of reason and amour de soi.

[82] Rousseau 2002b, p. 166.
[83] Rousseau 1979, p. 41. See Delaney 2006, p. 85 for further discussion of this point.
[84] Judith Shklar highlights this point in her remark that "[Amour-propre] subjugates the self in response to opinion and creates a second self, which in turn subjugates other men to these prejudices" (Shklar 1989, p. 90).
[85] See Douglass 2015, p. 172 for further discussion of this point.
[86] Rousseau 1979, pp. 252–253; O'Hagan 1999, pp. 86–89.

Clearly, there is a tight connection between freedom (co-independence of the will) and virtue. On the one hand, the domain of civil freedom is the domain of aims that are consistent with virtue. On the other, moral freedom just is virtuousness. In conjunction, they establish a 'reign of virtue' in which individual wills are aligned with the General Will.[87] What does this tell us about the aims relative to which agency is valuable and the conditions under which this value is realized? The first thing to note is that for Rousseau, virtue is a constitutive, and indeed central, element of happiness - that is, man's highest good.[88] Thus, to say that freedom is co-independence of the will is to say first, that agency is valuable because (a) there exist aims that are consistent with virtue, and (b) it is possible to be virtuous. If either (a) or (b) were false, then happiness or contentment would be an impossibility. And second, the value of agency is realized when (c) the ability of persons to pursue aims that are consistent with virtue is protected from external interference, and (d) persons are virtuous. Their joint satisfaction realizes the value of agency because, in securing co-independence of the will, civilized man secures his highest good: happiness.

To sum up, Rousseau's conception of freedom as co-independence of the will reflects the view that agency is valuable in virtue of aims that are conducive to happiness or contentment, but that the conditions under which this value is realized are context-sensitive. It is the mutability of human nature that explains this latter feature of his account, and why co-independence of the will is instantiated in seemingly contradictory ways across contexts. Natural freedom on the one hand, and civil and moral freedom on the other, do not describe different kinds of freedom per se, but rather what the very same kind of freedom looks like for two different kinds of creatures. In this light, the surface-level diversity of Rousseau's view of freedom is revealed to mask a deeper unity.

2.3.3 Criteria for APT revisited

In this section I have argued for a particular thesis about the value-dependence of conceptions of freedom: specifications of the ranges of possible value of each of the three variables of the triadic relation jointly specify the conditions under which the value of agency is realized. This is borne out by careful analysis of the kinds of considerations that Hobbes and Rousseau appeal to in the defence of their conceptions of freedom. Certainly, this does not defini-

[87] Dent 2005, p. 76.

[88] On the connection between virtue and happiness see ibid., p. 79 and Delaney 2006, p. 98.

tively establish that all conceptions of moral and political freedom appeal to these same kinds of considerations. However, given the magnitude of the differences between Hobbes' and Rousseau's accounts, we at least have good reason to suppose that my thesis holds more generally, including with respect to the myriad liberal conceptions of freedom that draw in different ways and to different degrees on the kinds of considerations that appear in both of the accounts examined here.

These findings give us a better idea of what a compelling case for APT must do. Recall that if FLP rules out transformative manipulation in principle, then it must be possible to demonstrate that

1. Transformative manipulation interferes with any plausible specification of liberal freedom

2. There are no sufficiently weighty reasons that could overcome the presumption against this kind of interference

In light of the analysis of freedom in this section, satisfying these criteria means demonstrating that from the perspective of any plausible conception of liberal political morality, such policy means necessarily interfere with conditions that realize the value of agency; and, furthermore, that there is something unacceptable about interfering in such a manner with realizing the value of agency. As we will see in Chapter 3, there are several candidate arguments for the latter claim. For now, it is enough that the desiderata for a compelling case for APT are clear.

Chapter 3

The Case Against Transformative Manipulation

FLP prohibits the state from undermining conditions that realize the value of agency unless there are sufficiently weighty reasons for doing so. Whether a given set of policy means is consistent with FLP depends, first, on one's conception of freedom, and second, the specification of the constitutive features of the means, i.e. instrument, method, mode, and content.[1] If the policy means do satisfy FLP, then one of the following conditions obtains: (a) all else being equal, none of its features (individually or in combination) undermine freedom; or (b) at least one of its features undermines freedom all else being equal, but its doing so is justified all things considered.

From a liberal perspective, it is difficult to escape the *prima facie* intuition that policies utilizing transformative manipulation never satisfy either (a) or (b). If this is correct, then FLP entails APT. My first aim in this chapter is to examine the case for APT. In §3.1 I argue that policies utilizing transformative manipulation undermine any plausible conception of liberal freedom, and so cannot escape APT by appealing to (a). There is a basic sense of autonomy that is embedded within a set of political commitments that any plausible account of liberalism endorses. Though theorists differ on *why* persons have an interest in the state's refraining from interference with basic autonomy, they do agree *that* such an interest exists. It is therefore constitutive of any plausible conception of liberal freedom. Transformative manipulation necessarily undermines autonomy in this sense, and therefore liberal freedom in general. In §3.2 I attempt to identify the strongest possible argument for the claim that policies utilizing transformative manipulation

[1] See §1.1.2

cannot escape APT by appealing to (b). I argue that the most compelling foundation for this position is a principle of respect for persons as valuable in themselves that grounds an inviolable duty of respect which constrains state action.

My second aim in this chapter is to challenge the respect-based argument against APT in a way that creates conceptual space for cases in which the use of transformative manipulation is justified on grounds of respect for persons. In §3.3 I argue that the degree to which rational persons are reasonably able to exercise basic autonomy is sensitive to political conditions. Part of what the duty of respect for persons demands of the state is that it maintain political conditions under which rational persons are mutually able to exercise this capacity to the greatest degree possible, even if they choose not to. If we assume that there are cases in which the use of transformative manipulation is necessary to maintain such conditions, then doing so is part of what the duty of respect for persons demands of the state, and as such, APT is mistaken. This does not establish that any such cases actually exist in practice. My aim in this chapter is simply to demonstrate that if such cases are conceptual possibilities, then liberal political morality does not rule out the use of transformative manipulation *in-principle*.

3.1 Transformative Manipulation and Liberal Freedom

In Chapter 2 I argued that conceptions of freedom designate conditions whose satisfaction realizes the value of agency. The liberal tradition contains a variety of views on what these conditions are, and so there is room for disagreement amongst liberals about when/if different policy means genuinely undermine freedom. For example, do coercive laws restrict liberty in ways that matter for freedom, or do they create the conditions that make the kind of liberty we care about possible?[2] Can information campaigns undermine individual autonomy? What about nudge policies that go about manipulating persons through choice architecture? Most liberals, I think, would agree that there is room within the liberal tradition for opposing answers to these questions. Transformative manipulation seems to be different, however. Intuitively, it is hard to see how policies utilizing this combination of method and mode could be consistent with *any* conception of liberal freedom. In the interest of addressing the strongest possible case for APT, my aim in this

[2] This question of course touches on the Lockean distinction between licence and liberty.

section is to defend this intuition.

I begin with the claim that there is a notion of autonomy that is constitutive of any plausible conception of liberal freedom. As many commentators have observed, liberalism is a broad philosophical tradition rather than a determinate political morality.[3] There are, however, a basic set of political commitments that liberals as a general rule agree upon:[4]

(a) Persons are free and equal from the political point of view; free as rational agents capable of practical reasoning, with plans and projects for their own life, and a capacity to understand and respond to moral reasons; equal as sharing the same fundamental moral status

(b) All persons have certain basic and equal rights such as freedom of thought, conscience, expression, and association, as well as rights associated with the proper functioning of democratic systems, bodily integrity, and private property

(c) The protection of these rights and liberties is one of the main functions of any legitimate state

(d) Even if these rights are defeasible, they have a certain priority in political reasoning, and are not easily defeated by conflicting considerations

Of significance here is the link in (a) between freedom and autonomy. For plans and projects to be our own in a meaningful sense, we must not only regard them as expressions of what we believe to be valuable in life, but we must also be sufficiently able to critically reflect on and revise our beliefs about the good. Similarly, the ability to understand and respond to moral reasons requires being sufficiently able to critically reflect on and revise principles of right that we take to be authoritative in our relations with others. Each of these involves a basic kind of reflexive agency, namely, the capacity to critically reflect on and revise one's normative commitments. I will refer to this capacity as 'basic autonomy'.[5]

Liberal states regard persons as having a weighty interest in the absence of state interference with basic autonomy. Of course, theorists disagree on what grounds this interest, the most fundamental cleavage being between political and comprehensive liberals. For political liberals, the interest is grounded in the challenge of maintaining a mutually acceptable political order against a

[3] See Waldron 2004 and Ryan 2007 for illustrative discussions.

[4] I draw these from Quong 2011, pp. 14–15 in an abridged form.

[5] This should not be confused with a *state* of autonomy, which results from the more or less effective exercise of the capacity.

background of reasonable value pluralism.[6] Reasonable persons within free societies often come to quite different conclusions on fundamental questions of value. The *raise d'etre* of public institutions is to secure conditions within which reasonable people can pursue their chosen way of life on mutually acceptable terms. Their popular legitimacy therefore depends on remaining as agnostic as possible on questions of value that are subject to reasonable disagreement. This requires respecting the basic autonomy of persons in their capacity as citizens within the public sphere, while also refraining from promoting specific doctrines or ways of life (autonomy-affirming or otherwise) in the private sphere. It undermines freedom when it fails to do so *because* this weakens the mutual acceptability, and therefore stability, of the political order.[7]

In contrast, comprehensive liberals typically ground the interest in claims about the value of basic autonomy itself.[8] Raz, for example, argues that a valuable life is one that is shaped by the autonomous choice of objectively worthwhile ends - in other words, that basic autonomy has constitutive value.[9] In a different vein, Hurka argues that autonomy is intrinsically valuable because it realizes an ideal of agency that aims at a particular kind of causal efficacy with respect to the world.[10] For views like these, interference with basic autonomy undermines freedom when it disrupts the realization of this value. Conversely, interferences that facilitate the realization of the value of basic autonomy enhance freedom - it is no coincidence that most comprehensive liberals endorse some form of perfectionism that permits states to promote autonomy-affirming conditions or ways of life over alternatives.

What this comparison illustrates is that disagreement between liberals about the nature of the interest in basic autonomy affects *how* it figures into a conception of freedom, but not *that* it does. Basic autonomy is constitutive of any plausible conception of liberal freedom because it is built into a set of political commitments that liberals as a rule agree on. The reason that liberals of all stripes can agree that transformative manipulation undermines freedom is that, from the perspective of both political and comprehensive liberals, such policies are incompatible with the interest persons have in basic autonomy. Recall that, roughly speaking, policies utilizing transformative manipulation attempt to alter a target population's normative commitments

[6] See Larmore 1999, Rawls 2005, Quong 2011, and Nussbaum 2011 for influential discussions of political liberalism.

[7] I return to the topic of stability in Chapter 4.

[8] See Gaus 2004 for a systematic overview of comprehensive liberalisms.

[9] Raz 1986.

[10] Hurka 1987.

in a manner that involves intentionally concealing these efforts (or the reasons for them) from the affected persons. Suppose that the state wants to improve population health by inducing a preference for certain healthy lifestyle choices. In an effort to accomplish this, they surreptitiously fund trusted non-governmental organizations to stigmatize unhealthy choices and valorize healthy ones. As a purely descriptive matter, the manipulative dimension interferes with basic autonomy because it intentionally deprives the target population of information that could play a role in their deliberations in response to the messaging. In contrast, the transformative dimension interferes with basic autonomy in the sense that it alters the course of its future exercise. Exercising basic autonomy involves reflective consideration of whether one endorses certain normative commitments as legitimate sources of reasons in one's deliberations. But the foundation for any these judgments can only be our other normative commitments. Therefore, anything that changes the composition of our commitments can (and given enough time invariably does) affect the outcome of subsequent exercises of basic autonomy.

From the perspective of the political liberal, the transformative dimension necessarily conflicts with the interest in basic autonomy. The state's attempting to affect a change to a population's normative commitments invariably means taking a stand on matters of reasonable disagreement. Given the practical reality of reasonable value pluralism, this can only erode the mutual acceptability of a liberal political order, and therefore its stability. Such policies are therefore a threat to freedom. Conversely, it's not obvious that the same can be said of manipulation. Policies that use conservative manipulation - for example, the kinds of nudge policies that Thaler and Sunstein discuss - can be designed to help people overcome internal barriers to doing what they already see themselves as having most reason to do.[11] This does not seem to involve taking a stand on matters of reasonable disagreement, or otherwise undermine the stability of a mutually acceptable political order amongst reasonable persons. The picture is reversed for the comprehensive liberal. There is no in-principle clash between the transformative dimension and the interest in basic autonomy here. If basic autonomy is non-instrumentally valuable, then *harnessing* it to affect a change to an agent's normative commitments - e.g. by using transformative persuasion - need not disrupt the realization of its value. Indeed, such policies can enhance rather than undermine freedom. The manipulative dimension, on the other hand, does seem to fundamentally clash with an interest in basic autonomy as non-instrumentally valuable. It's not clear how policies that aim

[11] Thaler and Sunstein 2008.

to surreptitiously distort or circumvent an agent's ability to exercise basic autonomy could do anything but disrupt the realization of this value on any interpretation.

These considerations do not amount to a proof that transformative manipulation necessarily undermines any plausible conception of liberal freedom, but they do strongly support the intuition that this is so. It would be a very strange conception of liberal freedom that was not undermined by at least one dimension of these policy means in-principle. I therefore leave aside any attempt to show that transformative manipulation can escape APT by arguing that such policies are consistent with freedom under certain conditions. In the next section I examine the case for the second premise of APT, namely that the manner in which policies utilizing transformative manipulation undermine freedom cannot be justified within a liberal order.

3.2 The Case Against Transformative Manipulation

Very little can be inferred about the permissibility of state action simply based on the fact that it undermines freedom. Most obviously, it matters *why* the action is undertaken. Objections to incarcerating non-violent drug offenders typically focus on the justification for this action, not the legitimacy of incarceration as such. Similarly, those who oppose abortion bans do so because they think the balance of reasons favours access to legal abortion services, not because they dispute the legislative authority of the state in general. Because states are afforded vast powers, most debates about the permissibility of state action pertain to 'why' questions. However, there is another, stronger kind of objection which concerns the 'how' rather than 'why'. The worry here is not that otherwise legitimate policy means are used unjustly, but rather that *any* use of these means is unjust. Many capital punishment abolitionists, for example, object to the practice of state-sanctioned execution *as such* - the state's motivation for sentencing a particular person to death is simply irrelevant. The second premise of APT is essentially a 'how' objection to policies that utilize transformative manipulation - it denies that there are any reasons that would allow such policies to overcome the liberal presumption against interference with individual freedom. In this section I consider four strategies for supporting this thesis, and argue that the strongest is one based on a principle of respect for persons.

3.2.1 The argument from ineffectiveness

One possible motivation for an absolute prohibition on transformative manipulation is a general scepticism about the efficacy of state interference with basic autonomy. Stephen Wall expresses something like this worry when he asserts that

> ...the state is generally not an effective instrument for cultivating mental capacities and virtues. When the state attempts to improve individuals' psychologies or remove intrapersonal barriers, it is more likely to do more harm than good. The state that attempts to make its subjects masters of themselves will likely just end up oppressing them[12]

There is no doubt something to this worry. The difficulty of formulating sound policy, paired with the unavoidable messiness of implementation, provides myriad opportunities for things to go wrong. A policy to promote reasonable tolerance as a value might end up fostering a zealous ideal that ultimately undermines the freedom of those who are perceived to fall short of it. Or, attempts to promote tolerance through manipulative means could push people towards intolerance if they regard such efforts as insulting. As with any kind of policy, failures and unintended consequences are distinct possibilities. Nevertheless, efficacy-related concerns are an unconvincing basis for an absolute prohibition on transformative manipulation. For one, the objection rests on an empirical claim about the efficacy of certain kinds of state action, and its not clear that it is supported by the facts. Indeed, in a footnote to the above remark, Wall recognizes (in a limited fashion) that certain kinds of state action can effectively influence a person's normative commitments:

> There are a few things that the state can effectively do [to cultivate mental capacities and virtues]. For example, it can do its best to ensure that all children receive and adequate education. But even here there are serious limits to what the state can do.[13]

Furthermore, even if the facts do support Wall's initial claim, they almost certainly reflect contingent epistemic and technological limitations on state action rather than insurmountable barriers, e.g. the state's access to reliable data on public attitudes, scientific understanding of psychological processes, etc. If this is correct, then sufficient advances in these areas would presumably make transformative manipulation a legitimate part of the state's

[12] Wall 2003, p. 308.
[13] Ibid., 308n3.

toolkit. Unless a good argument can be made as to why such advances are implausible, the ineffectiveness argument does not support the claim that transformative manipulation is impermissible *in-principle*.

3.2.2 The argument from instrumental value

Another possibility is that transformative manipulation is impermissible because the manner in which it interferes with basic autonomy hinders the realization of other important goods. Mill, for example, argues that both individual flourishing and social progress require that "...free scope should be given to varieties of character...the worth of different modes of life should be proved practically, when any one thinks fit to try them".[14] This claim is motivated in part by the concern that third parties (including the state) are liable to err in their judgments about what is good for others. Their opinions may be shaped by overly-narrow experience or misinterpretation of what experience teaches, or mistaken beliefs about the generalizability of the lessons they have drawn from experience. As such, it is each individual that is the best judge of "...what part of recorded experience is properly applicable to his own circumstances and character".[15] Furthermore, imposing beliefs about the good on someone degrades their capacity to make such judgments, rendering them 'inert and torpid' and so less able to recognize and desire that which their flourishing consists in.[16,17]

Some liberal anti-perfectionists adopt an instrumentalist justification for protecting basic autonomy by making it a condition for establishing and maintaining the legitimacy of political institutions. John Christman, for example, argues that political institutions are legitimate only when the are non-alienating in the sense that their actions

> ...can be seen as harmonizing with our own judgments, our perspectives about what is valuable to pursue given the fact that we live among people with contrasting values, and who (like us) are the products of the contingencies of history...[18]

[14] Mill 2003, p. 122.
[15] Ibid., p. 123.
[16] "...to conform to custom, merely as custom, does not educate or develope in him any of the qualities which are the distinctive endowment of a human being...He who does anything because it is the custom...gains no practice either in discerning or in desiring what is best. The mental and moral, like the muscular powers, are improved only by being used" (ibid., pp. 123–4)
[17] Mill also provides non-instrumental reasons for protecting basic autonomy which I cover below.
[18] Christman 2009, p. 225.

Democratic processes are central to achieving this goal - they provide mechanisms for people to express their respective normative commitments such that political institutions are prevented from becoming dominated by a subset of interests. Crucially, the integrity of these processes is predicated on the protection of basic autonomy, for

> Democratic deliberation...requires participants' abilities to reflectively endorse, indeed publicly defend, the points of view, values, interests, and opinions that are the inputs to such deliberative processes (the "outputs" of which are social principles and policies)[19]

Policies that utilize transformative manipulation appear to threaten the integrity of these processes by raising questions about whether citizens are expressing *their* normative commitments in a robust sense, or rather those of policy-makers or special interests. If the latter, then it would seem that these processes are so much political window-dressing, which would undermine the perceived legitimacy of public institutions themselves.

Instrumentalist arguments against transformative manipulation can support a strong *pro tanto* presumption against its use, but invariably fail to support the stronger claim of an absolute prohibition. One source of error is a tacit assumption about the immutability of the background conditions that make the protection of basic autonomy necessary as a means to achieving other goods. For example, Mill's claims about the sources of error in third-party judgments about what is good for others only support an absolute prohibition on transformative manipulation if they cannot be meaningfully mitigated. But is it a necessary truth that such judgements can only reflect overly-narrow personal experience, or mistaken beliefs about whether the lessons one draws for oneself are applicable to others? To be fair, Mill certainly did not seem to think so. He is careful to claim only that one's personal experience "*may* be too narrow" and that one's interpretation of experience "*may* be correct, but unsuitable to [another]" [emphasis added].[20] In this he is undoubtedly correct, but it's not clear that these sources of error cannot be effectively mitigated, if only by paying careful attention to signs of their influence in our judgments.

Another source of error is the inference that because the protection of basic autonomy in some respect is necessary to achieve other goods, *any* interference with it must hinder our achievement of these goods. Consider Mill's worry about the stunting effect of interference with basic autonomy

[19] Christman 2009, p. 226.
[20] Mill 2003, p. 123.

- certainly, this is plausible under conditions of pervasive interference. But suppose the state utilizes transformative manipulation only to bring individuals to value the exercise of basic autonomy as a means for determining what is best in life. Far from stunting this capacity, such a policy would presumably strengthen it, and thus make a positive contribution to individual flourishing and social progress. A similar argument can be made with respect to Christman's worry about the impact of interference with basic autonomy on the perceived legitimacy of political institutions. Undoubtedly, using transformative manipulation to influence the *outputs* of democratic processes could lead some people to question the democratic legitimacy of political institutions. But what about a policy to promote participation in democratic processes by bringing citizens to regard it as an important duty? Presumably, the more engaged people are in these processes the more responsive political institutions are likely to be to the actual distribution of interests in the population, and therefore the stronger their perceived legitimacy. Of course, one may object that the use of transformative manipulation in both cases is liable to backfire or otherwise go wrong, but this is to fall back on the efficacy argument.

3.2.3 The argument from intrinsic value

Another possibility is that transformative manipulation is prohibited because basic autonomy has profound intrinsic value, i.e. it has positive value that cannot be reduced to the value of anything else. As an illustration, consider Hurka's view that autonomy realizes an ideal of agency as being a causally efficacious actor in the world.[21] He asks us to consider two teachers, one who chose her profession from a set of viable options - e.g. to become a plumber, lawyer, etc - and another who did so because it was her only option. All else being equal, the first teacher is more autonomous than the second because she not only chose to become a teacher, but also chose to *not* become a plumber or lawyer. That she is not one of these other things is explained by facts internal to her. The same cannot be said of the second teacher. Though at some level she chose her profession, she did not choose to not be a plumber or lawyer. That she is not one of these other things cannot be explained by facts internal to her alone. Hurka argues that the first, more autonomous teacher better satisfies the ideal of agency because she is "...responsible for more facts about her life, and thus is more expansively an agent".[22] This logic

[21] Hurka 1987. His claim is not that autonomy is a *means* to realizing the ideal of agency - it *is* the realization of this ideal.

[22] Ibid., p. 143.

applies to basic autonomy if we assume that our normative commitments are included in the facts about our lives that we may be responsible for. In this case, the more robust our ability to exercise basic autonomy - e.g. the more insight we have into our normative commitments and the more possible alternatives we are exposed to - the greater the degree to which the ideal of agency is realized.

Of course, if basic autonomy is just one of many things that are intrinsically valuable, then it's not clear why transformative manipulation couldn't be justified when it facilitates the realization of sufficient amounts of these other goods. But what if no amount of other goods can compensate for the loss of basic autonomy? Suppose that it is *unconditionally* valuable, i.e. its intrinsic value is such that more of it is always better, both in terms of increasing quantities and the choice between it and other sources of value.[23] Or that it is incommensurable or incomparable with other intrinsically valuable things, e.g. basic autonomy is special in a way that precludes its substitution by other goods without a net loss of realized value; or there is no common scale that would make measuring trade-offs between basic autonomy and other goods possible, etc.[24] While these possibilities preclude compensation via the realization of other goods, neither precludes compensation via basic autonomy itself. Consider the use of transformative manipulation to induce a belief in the value of critical reflection on one's normative commitments in a population whose social norms discourage this. Such a policy interferes with basic autonomy for the reasons discussed in §3.1; but the end result is that the members of the target population possess a more robust capacity to critically reflect on and revise their normative commitments. If this capacity is unconditionally valuable (and so more is always better), then the use of transformative manipulation in this case seems to be perfectly justifiable. And because this judgment only involves a comparison between amounts/degrees of basic autonomy, issues of incommensurability and incomparability do not arise.

It is not clear how arguments that appeal to the intrinsic value of basic autonomy can overcome the compensation problem. I will not pursue this line of questioning any further, however, for there exists a more promising strategy for establishing the second premise of APT.

[23] This definition is a paraphrase of Carter's 1999, p. 39.
[24] See Chang 2013 for a succinct discussion of these and related ideas.

3.2.4 The argument from respect for persons

The previous arguments attempt to motivate the second premise of APT by appeal to obligations that states have to protect or promote the welfare of its citizens. But as we've seen, it's not obvious that transformative manipulation conflicts with these requirements in all circumstances. There is another strategy available to the defender of APT, however, one that rests not on what is valuable *for* persons, but rather on the value *of* persons - that is, on duties of respect. In outline, the argument is this: the state is under a duty of respect for persons *as* persons ('respect for persons'); transformative manipulation is necessarily disrespectful of persons; therefore, utilizing transformative manipulation as a policy means is impermissible in-principle.

Liberal political commitments express two fundamental claims about the value of persons. First, persons are (or are to be regarded as) valuable in themselves - that is, irrespective of demographic characteristics (ethnicity, nationality, sex, gender, etc), beliefs, actions, or circumstances. 'Value-in-itself' is a notoriously tricky idea, but for the present purposes it is sufficient to understand it as a kind of intrinsic value whose realization does not depend on its bearer being good for anything/anyone else.[25] And second, the value of persons outweighs, or is lexically prior to, the value of anything else. Actions that affect persons can be justified only on the basis of their interests.

Value and respect are closely connected.[26] It is the very nature of value that it constitutes (or provides) a reason to conduct oneself towards its bearer (x) in a manner that is consistent with the realization of this value. Not all of these are reasons of respect, however. As Raz observes, reasons to engage with x such that its value is in fact realized apply only to those for whom doing so is good.[27] That aesthetic engagement with the sculpture enriches the lives of many people means that it has value, but if it does not similarly enrich *my* life then its value does not give me a reason to engage with it in this way. And yet because it has value, I still have reason to respect the sculpture. Reasons of respect motivate a different sort of engagement: first, when we think of something of value, to adopt psychological attitudes or beliefs about it that are appropriate to its value; and second, to conduct oneself in ways that are consistent with the preservation of valuable things, including not

[25] See Raz 2004, pp. 151–152 for a more detailed explication of this idea. For another recent treatment of value-in-itself, including its Kantian origins, see C. M. Korsgaard 2021.

[26] I leave aside conceptions of respect that do not have any moral connotations, e.g. respecting the power of the ocean.

[27] Raz 2004, pp. 162–163.

destroying them.[28] In other words, reasons of respect are reasons to engage with x in a manner that is consistent with the ongoing *possibility* of its value being realized, even if we do not ever intend to engage with it in a way that actually realizes its value.[29]

The content of reasons of respect depends on the nature of the object and the kind of value it has. As noted above, that persons are valuable in themselves means that realizing their (intrinsic) value as persons does not depend on their being good for anyone/anything else. Instead, it is realized when they are able to live their lives *as* persons. If we accept that the capacity to critically reflect on and revise one's beliefs about what matters is a constitutive feature of personhood, then the absence of interference with this capacity is a necessary condition for realizing one's value as a person.[30] Respecting persons therefore involves, first, adopting attitudes/beliefs that acknowledge the constitutive value of basic autonomy by virtue of its connection to personhood;[31] and second, conducting oneself in a manner that is consistent with the preservation of persons' basic autonomy. To satisfy these conditions is to confer authority on persons in one's relations to them, to accept that certain facts about them *as* persons impose constraints on how they are to be engaged with. What this describes is a form of *recognition* respect.[32] That the state is under a perfect - rather than merely *pro tanto* - duty of respect for persons in this sense follows from the fact that, since persons are valuable in themselves and this value is weightier than the value of anything else, reasons of respect for persons are overriding moral reasons.

The preceding considerations clarify an important aspect of what respect for persons involves and why the state is under a duty to conform to it. In doing so, they also highlight why transformative manipulation is impermissible in principle. Since policy means of this kind interfere with basic autonomy, they also impair the realization of the value of persons as such. By failing to engage with them in a manner that reflects their value, such policies fail to engage with them *as* persons - that is, they are fundamentally disrespectful of persons. And since the value of persons grounds a perfect duty to com-

[28] Raz 2004, pp. 161–162. See also 1998, p. 104.

[29] Raz 2004, p. 167.

[30] This would appear to imply that not all human beings are persons and therefore deserving of respect, e.g. children, those with profound developmental disabilities, etc, or that no animals are deserving of respect. There are a variety of plausible ways of meeting these challenges, but I will leave these issues aside for the purposes of this discussion. See Wood 2008, pp. 95–105 for an insightful discussion of this topic.

[31] Something has constitutive value when it is "...definitive of a larger complex that is itself valued" (Dworkin 1988, p. 80).

[32] Darwall 1977, 2006.

ply with reasons of respect, there are no circumstances in which the use of transformative manipulation can be justified.

The argument from respect for persons has two major advantages over the preceding arguments for the second premise of APT. First, it is not vulnerable to the objections levelled against them. That transformative manipulation is disrespectful of persons, and that the state is under a perfect duty of respect for persons is not contingent on any epistemic or technological facts, or our ability to overcome our own fallibility in making judgements about the good of others. Rather, these conclusions follow from the value of persons and the nature of respect themselves. And since engaging with persons in a manner that appropriately reflects their value is what respecting persons consists in, there are no circumstances in which transformative manipulation could be a *means* for respecting persons. And finally, because persons are intrinsically valuable in ways that supersede anything else of intrinsic value, it is not possible to justify the use of transformative manipulation by realizing any amount of the latter.

The second advantage of the argument from respect for persons is that it makes sense of the intuition that transformative manipulation is morally objectionable whether or not it succeeds in a given case. As noted above, respecting persons means adopting attitudes or beliefs that appropriately reflect the value of persons *and* conducting oneself in a manner that is consistent with the ongoing possibility of realizing this value. A successful application of transformative manipulation clearly violates both conditions. A failed one, meanwhile, merely violates the first, but this is enough to make it disrespectful of persons. In contrast, the argument from inefficacy can account for why failed attempts are bad, but not why successful ones are. Conversely, the arguments from instrumental and intrinsic value make a plausible case for why successful attempts are bad, but struggle to explain why failed ones are.

Despite these advantages, the argument from respect for persons faces an objection that appears to undermine the claim that it supports the second premise of APT. Richard Dean has challenged the plausibility of grounding a principle of equal respect for persons in the possession of a capacity of any kind (e.g basic autonomy).[33] The core of his argument is this: if respect for persons is a fundamental moral principle that applies to everyone in equal measure, then it must be on the basis of some property that (a) all persons possess in equal measure, and (b) is morally significant. Many philosophers

[33] Dean 2021. As he notes, similar objections have been raised in the past by Cranor 1982 and Neumann 2004.

have argued that this property is a capacity of some kind, e.g. rationality, morality, etc. However, the plausibility of these proposals trades on an ambiguity about what is meant by 'capacity'. On the one hand, it can refer to an unrealized potential, e.g. "She has the capacity to become an excellent philosopher if she applies herself, but who knows if she'll ever realize that potential".[34] On the other, it can refer to an ability that has been more fully realized, e.g. "She certainly has the capacity to write a lot of excellent papers, who knows where she finds the time".[35] To satisfy (a), the proposed capacity must be an unrealized potential, for people vary in whether and how well they exercise their capacities. If 'capacity' refers to a realized ability, then many people would not deserve respect. However, it doesn't seem plausible that an unrealized potential can satisfy (b). Dean notes that many writers argue that the capacity for rationality has deep moral significance because it "...helps us live reasonably decent lives and engage in successful cooperative ventures".[36] Interpreted as an unrealized potential, it's not clear how rationality could serve these ends. Surely the mere *potential* to act rationally does not help us live decent lives together - rather, it is the exercise of this capacity that does so. As such, rationality has moral significance only if we interpret it as a realized ability.[37] But as we saw above, respect for persons cannot be based on a realized ability, for such an account will inevitably fail to satisfy (a).

Dean's argument appears to create the following dilemma for the defender of APT: if basic autonomy is interpreted as an unrealized potential then it's not clear why it has sufficient moral significance to ground respect for persons, and therefore a prohibition on transformative manipulation. On the other hand, if it is interpreted as a realized ability, then the argument from respect for persons only establishes the in-principle impermissibility of using transformative manipulation on those who do in fact exercise their capacity for basic autonomy. In either case, then, the argument from respect for persons fails to establish the second premise of APT.

This dilemma is illusory, however. To see why, it is useful to consider

[34] Dean 2021, p. 142.
[35] Ibid., p. 142.
[36] Ibid., p. 144.
[37] Dean summarizes the objection in the following passage: "The claim that some unrealized capacity or potential demands profound respect, in fact the most profound respect, in and of itself is inconsistent with the way that we usually think of the reactions that are appropriate to an unrealized capacity or potential. In general, the treatment demanded by a mere potential is intrinsically tied to the eventual realization of the potential, the development of the actual ability or trait" (ibid., p. 145).

some examples of the views that Dean addresses:

> ...the possession of humanity [i.e. the power of rational choice] and the capacity for the good will, whether or not that capacity is realized, is enough to establish a claim on being treated as an unconditional end.[38]

> ...respecting the value of human (rational) life requires us to treat rational creatures only in ways that would be allowed by principles that they could not reasonably reject insofar as they, too, were seeking principles of mutual governance.[39]

What is significant is that in both cases, the capacity that defines persons *as* persons and the capacity that explains why persons warrant respect are one and the same. It is this fact that leads Dean to challenge the idea that the possession of a mere capacity could be morally significant enough to support a duty of equal respect for persons. But there is no reason that the specified capacity must play this dual role in an account of respect for persons. Indeed, the respect-based argument for the second premise of APT is a clear illustration of this. Basic autonomy is a constitutive feature of persons, and therefore figures into the description of what respect for persons involves. But persons do not warrant respect *because* they possess this capacity - it is because persons are valuable in themselves, i.e. the realization of their value does not depend on their being good for anything/anyone else. Dean's objection fails against views of this kind because they invert the relationship between the value of persons and the value of the capacity that defines persons: persons do not warrant respect because they possess a morally significant capacity - rather, the capacity is morally significant because persons warrant respect.[40]

My aim in this section was to identify the strongest case against the permissibility of transformative manipulation as policy means. I have argued

[38] C. Korsgaard 1996, pp. 123–124.

[39] Scanlon 1998, p. 106.

[40] Sarah Buss touches on a similar point when she argues against the view that the value of autonomy *as such* explains why undermining it is wrong: "When deceit and manipulation are morally impermissible, this is because they prevent someone from being (sufficiently) in touch with reality, or from relating to the manipulator as an equal, or because they reflect a lack of concern for the well-being and happiness of another human being...to treat someone in this way is to do something that a reasonable agent cannot autonomously endorse. In this sense, it is to treat the person's capacity for rational, autonomous choice with contempt. This is a bad thing. But there is nothing about the nature of autonomous agency itself that explains why it is bad" (Buss 2005, p. 234).

that a principle of respect for persons as valuable in themselves provides the most promising foundation in this regard. Of course, it is possible that I have overlooked something. Perhaps there are stronger arguments that are based on something other than respect for persons, or perhaps on a different analysis of the nature of respect. Even if there are, however, any case for the permissibility of transformative manipulation will still have to defuse the argument from respect for persons presented here. In lieu of any stronger objections, I therefore focus my efforts in the next section on showing why this argument does not show that transformative manipulation is incompatible with respect for persons, and consequently, why APT in its strongest form must be rejected.

3.3 Making Room for Transformative Manipulation

Recall that APT rests on two premises: transformative manipulation undermines any plausible conception of liberal freedom by virtue of its impact on basic autonomy; and undermining freedom in this way is impermissible in-principle because it is inherently disrespectful of persons. In this section I sketch an argument against APT that is fleshed out in subsequent chapters. In particular, I argue that the degree to which rational persons are reasonably able to exercise basic autonomy is sensitive to political conditions.[41] Part of what the duty of respect for persons demands of the state is that it maintain political conditions under which persons are able to exercise this capacity to the greatest degree possible (even if they choose not to). Therefore, if there are circumstances in which the use of transformative manipulation is necessary to maintain such conditions, then its use is not inherently disrespectful of persons. As noted in the introduction to this chapter, this argument does not establish that such circumstances exist, for it may not be the case that transformative manipulation *is* ever necessary to secure the requisite conditions. My aim here is simply to demonstrate that there is conceptual space within liberal political morality for its justification as policy means.

[41] By 'rational' I just mean that an individual has the ability to set and revise ends, and more or less effectively pursue them.

3.3.1 The effect of political conditions on basic autonomy

I have described basic autonomy as a constitutive capacity of persons to both critically reflect on and revise their normative commitments. We can understand critical reflection as a process by which persons arrive at higher-order pro or con attitudes towards their first-order normative attitudes.[42] For example, P might come to disapprove of his desire to eat meat because it conflicts with the value he places on animal welfare. Revision takes things a step further, being the process by which we come to endorse or reject certain of our normative attitudes based on the outcome of critical reflection. But what exactly is the difference between having a pro or con attitude towards a normative attitude α and endorsing or rejecting α? Michael Bratman notes that it is a characteristic feature of human agency that

> We do not simply act from moment to moment. Instead, we settle on complex...future-directed plans of action, and these play basic roles in support of the organization and coordination of our activities over time.[43]

Amongst these plans are what he refers to as *self-governing policies*. These are, in essence, higher-order judgments about which of our first-order attitudes are legitimate considerations in practical reasoning, and their relative weight. For example, suppose that P is a soldier who has a desire for extrajudicial retribution against captured enemy forces, but because P values the rule of law, he refuses to countenance this desire as a legitimate consideration in deliberations about how to treat them. Such judgments about which first-order normative attitudes are and are not reason-giving reveal the presence of a self-governing policy rather than merely a higher-order pro- or con-attitude. What it means to endorse a first-order normative attitude α, then, is that one has a self-governing policy that counts α as reason-giving; conversely, rejecting α means having a self-governing policy against α as reason-giving.[44] What I have been referring to as normative commitments are effectively self-governing policies, It follows that revising one's norma-

[42] I am utilizing a hierarchical model in the vein of Frankfurt 1971 here, but my analysis could just as easily be cast in terms of other models of reflective endorsement.

[43] Bratman 2007, p. 26.

[44] In Bratman's terms "...we should understand an agent's endorsement of a desire in terms, roughly, of a self-governing policy in favor of the agent's treatment of that desire as providing a justifying reason in motivationally efficacious practical reason" (ibid., p. 39).

tive commitments is a process of rejecting or endorsing certain first-order normative attitudes as reason-giving.

If persons' ability to exercise basic autonomy is sensitive to political conditions, then it is because the latter influences which higher-order attitudes they would likely form and/or which of these higher-order attitudes are likely to be converted into normative commitments upon the exercise of this capacity - that is, upon the exercise of both critical self-reflection and revision. It seems plausible that political conditions do affect the degree to which rational persons can meaningfully undertake critical reflection. Under a totalitarian regime, for instance, internalizing state-sanctioned first-order normative attitudes - say, valuing conformity over individual initiative - can be a matter of survival. Given sufficiently extreme surveillance and repression, it seems plausible that one would refrain from, or at least greatly restrict, critical reflection on the relevant first-order normative attitudes. However, it's also possible that persons are just extremely careful about expressing the higher-order attitudes that arise from critical reflection under these conditions, e.g P loathes his desire to conform that the regime has cultivated in him, but never expresses this attitude for fear of punishment. For the sake of argument, then, let us assume that critical reflection is *not* sensitive to political conditions.

The same assumption cannot be made about revision. As noted above, normative commitments determine which first-order normative attitudes count as reason-giving and their weight in practical reasoning. Leaving aside extreme forms of akrasia, part of what it means for P to accept a first-order attitude α as a reason of weight ω to ϕ is that there are circumstances in which P would be *moved* by α to ϕ by in virtue of ω. In other words, P truly accepts α as a normative reason only if α is a potentially motivating reason for P. Now, a person's circumstances clearly influence which first-order attitudes they regard as reason-giving and their weight. These circumstances can include facts about ourselves - for example, because P acknowledges that she is an alcoholic, she strongly rejects her desire to drink as a consideration in favour of consuming alcohol - but more commonly our normative commitments are shaped by facts about the world, including the political conditions within which our lives play out.[45] There are at least three ways that political conditions can affect the degree to which rational persons are able to meaningfully revise their normative commitments.

[45] People can of course make choices in their private lives that influence the content of their normative commitments, e.g. joining a religious community, but this just illustrates that political conditions are not the *only* exogenous influence on their content.

First, state action can exert a strong influence on the weights people are reasonably able to assign to their normative commitments in practical reasoning. Suppose that P is motivated by a sense of civic duty to write popular articles exposing government corruption. In response, the government enacts a decree that permits the indefinite detention of those who are critical of the state. We can expect that for the purposes of deciding whether to continue writing, the relative weight of P's sense of civic duty will be greatly reduced in the wake of the decree. Furthermore, it is reasonable to expect that P's willingness to assign greater weight to his sense of civic duty will be greatly constrained by his awareness of the potentially severe costs of doing so.[46]

Second, state action can effectively restrict which first-order normative attitudes rational persons can reasonably consider as candidates for endorsement or rejection. For example, given sufficiently severe restrictions on free speech it is not unreasonable to expect that many people will, as a matter of indoctrination/internalization, come to reject any desire to speak freely as ever counting in favour of doing so. State action can therefore not only influence the weights persons assign to various first-order attitudes, but also the range of first-order attitudes that they can imagine assigning any weight to at all.

Third, state action can effectively extinguish the ability of persons to revise their normative commitments altogether. Orwell vividly describes this state of affairs in *Nineteen Eighty-Four* as one in which individuals "...had to live - did live, from habit that became instinct - in the assumption that every sound you made was overheard, and...every movement scrutinized".[47] To live by instinct in this sense is to have one's normative commitments so completely determined by outside influences that revising them in any meaningful sense becomes unthinkable. Orwell's depiction, though extreme, is no mere fantasy. Consider the following description of life in North Korea:

> If North Koreans paused to contemplate the obvious inconsistencies and lies in what they were told, they would find themselves in a dangerous place. They didn't have a choice. They couldn't flee their country, depose their leadership, speak out, or protest. In order to fit in, the average citizen had to discipline himself not to think too much.[48]

Perhaps it is impossible for a state to completely extinguish a person's ability

[46] There will always be exceptions of course. Some people will continue to criticize a tyrannical state even on penalty of death because they value personal freedom so strongly.
[47] Orwell 2021, p. 5.
[48] Demick 2010, p. 70.

to revise their normative commitments without destroying them as a person, but the testimony of those who have survived regimes like these demonstrates that it can be done for all intents and purposes.

3.3.2 The provisional argument against APT

A central part of what the duty of respect for persons requires of the state is that it secure and maintain political conditions under which rational persons are, to the greatest possible degree, reasonably able to revise their normative commitments. This includes refraining from actions that reduce the degree to which they are able to do so unless there is sufficient justification for doing so. As such, the state has at least a *prima facie* reason to forbear from implementing policies that produce any of the three effects discussed above. The question that confronts us is whether, in each case, it is anything more than a *prima facie* reason.

When it comes to state actions that restrict the degree to which rational persons are reasonably able to assign particular weights to their first-order normative attitudes, the reason must be a *pro tanto* one. Securing and maintaining political conditions within which rational persons have the greatest possible leeway to revise their weightings surely requires the enforcement of laws against, for example, murder, theft, etc. While such laws themselves restrict the degree to which rational persons are reasonably able to weigh various first-order normative attitudes, the magnitude of the restriction pales in comparison to what they would face under alternative political conditions, e.g. anarchy, despotism. Restrictions of this kind can therefore be justified if it is necessary to avoid the establishment of even more restrictive political conditions - indeed, doing so is an expression of respect, for it amounts to ensuring the ongoing realization of their value as persons.

Things are murkier in the case of state action that effectively restricts the ability of rational persons to regard (or disregard) some subset of their first-order normative attitudes as reason-giving altogether. If there exist scenarios in which such actions are necessary to preserve conditions that afford rational persons the greatest possible leeway to revise their normative commitments over time, then the *prima facie* reason is in fact a *pro tanto* one; conversely, if no such scenarios exist, then the reason is overriding. It is my aim in subsequent chapters to show that such scenarios do in fact exist, so for now let us provisionally assume that they do. Under this assumption, state action of this kind can be an expression of respect for persons and is therefore justifiable at least in-principle.

Finally, it seems clear that the *prima facie* reason against creating conditions that effectively extinguish persons' ability to revise their normative

commitments is an overriding one. Given the unconditional value persons and the role of basic autonomy in realizing this value, what could possibly justify state actions that have this effect? Perhaps there are cases (however remote) where such conditions are a necessary step to securing conditions in which rational persons have, to the greatest possible degree, the ability to revise their normative commitments. This comes perilously close to a contradiction, however; it looks a lot like burning the village in order to save it. Even if strictly speaking there is no logical contradiction here, as an empirical matter there is also little reason to believe that this degree of repression ever gives rise to political conditions that are conducive to the exercise of basic autonomy to any degree - repression typically begets repression. For practical purposes it is fair to conclude that state action that creates such conditions necessarily violates the principle of respect for persons and is therefore impermissible.

Nothing that I have said so far is incompatible with the second premise of APT. Certainly the proponent of this view would agree that state actions that effectively extinguish persons' ability to revise their normative commitments are impermissible. Similarly, they would agree that affecting persons in the other two ways is justified if it is necessary to preserve political conditions under which rational persons are, to the greatest possible degree, reasonably able to revise their normative commitments. What they specifically object to is the use of transformative manipulation to achieve this end. Clearly, there are ways to restrict the weights that rational persons assign to their first-order normative attitudes that do not require the use of transformative manipulation. Laws, for example, do so by appealing to their existing normative commitments, i.e. they are policies in the conservative mode. There are also ways to bring people to be effectively unable to regard (or disregard) some sub-set of their first-order normative attitudes as reason-giving that do not involve the use of transformative manipulation. For example, using rational persuasion to bring them to see that certain of their first-order normative attitudes are contrary to their most deeply held values. Policy means of these kinds restrict the degree to which rational persons are reasonably able to revise their normative commitments, but they can be designed to do so in ways that *engage* with this capacity, i.e. that engage with persons *as* persons. In an important sense, they (when properly designed) make persons party to the relevant restrictions, rather than passively subject to them. Transformative manipulation, on the other hand, involves hijacking or overriding persons' capacity to revise their normative commitments, and therefore fails to engage with them as persons. It is this fact that motivates APT. The point of contention, then, is not whether state action that effectively restricts the degree to which rational persons are reasonably able to

revise their normative commitments can be justifiable. Rather, it is whether doing so in a way that fails to engage with them as persons can be justified.

APT faces a serious problem here. It requires that the state restrict the ability of rational persons to revise their normative commitments if (a) it is necessary to preserve political conditions under which it is possible to exercise this capacity to the greatest degree possible, but only insofar as (b) the means for achieving this end engage with persons *as* persons. As discussed above, satisfying (a) is part of what the duty of respect demands of the state, while (b) is a respect-based constraint on the means of achieving (a). But what if the only plausible way to satisfy (a) is by utilizing transformative manipulation and therefore violating (b)? The defender of APT cannot give up (a) in order to save (b), for, this would mean that the state does not violate its duty of respect when it does nothing to prevent a slide into political conditions that profoundly restrict the degree to which rational persons are reasonably able to revise their normative commitments.

This cannot be the case since, as we've seen, the unconditional value of persons places the state under a perfect duty of respect to satisfy (a). But neither can they give up (b) in order to save (a), for this would mean either abandoning the requirement that the state is under an absolute duty of respect for persons, or accepting that respecting persons does not necessarily require engaging with them *as* persons. The first path is a non-starter, for if the state is not under a duty of respect for persons then its not clear why they have a duty to satisfy (a) at all. Taking the second path, however, means rejecting the second premise of APT, and therefore APT itself.

The only other option that the defender of APT has here is to deny that there *are* any cases in which the use of transformative manipulation is the only plausible means for achieving (a). Even if we accept this, however, it concedes something important. By transforming the second premise of APT into an empirical claim, it undermines the overall thesis that transformative manipulation is incompatible with liberal political morality *in-principle*. In its place we find a *pro tanto* presumption against its use, albeit a very strong one. Liberal political morality can therefore as a conceptual matter accommodate the use of transformative manipulation as policy means since doing so isn't *necessarily* disrespectful of persons, even if we are hard-pressed to identify real-world cases where its use would be an expression of respect for persons.

As discussed in §3.1, different strains of liberalism can accommodate the use of either transformative or manipulative means under certain circumstances, but few if any liberals would accept their use in tandem. This tells us something about how deep the conflict with liberal political morality is perceived to be. However, the strength of this intuition sits uneasily with the

findings here. We don't typically think that the state's manipulating persons into endorsing or rejecting certain first-order normative attitudes as merely *contingently* wrong. Rather, it seems to be antithetical to liberalism's basic orientation towards the nature of persons and their value. If the argument in this section is correct, then this is not so. Perhaps it is of some consolation to the defender of APT that if there are no real-world cases in which transformative manipulation is justified, then the conceptual and empirical readings of APT are extensionally equivalent. The practical price for conceding that liberal political morality does not rule out the use of transformative manipulation in-principle therefore appears at this stage to be rather small.

3.4 Conclusion

In this chapter I have examined whether the liberal presumption against state interference with individual freedom can accommodate the use of transformative manipulation as policy means. The conclusion that we have reached is that the presumption does not rule out such policies in-principle, though if we accept what I take to be the strongest argument against their use, it effectively does so in practice. If this were the end of the matter, then my argument merely identifies a conceptual quirk of liberal political morality that has no real consequences. I do not believe this to be the case, however. The claim that there are no plausible real-world cases in which the state's use of transformative manipulation can be justified is erroneous. In the following chapters I explore three such cases, each of which concerns persons who, by virtue of their normative commitments, represent a realistic threat to the stability of political conditions under which rational people are reasonably able to revise their normative commitments. If my arguments are sound, then we must reject both the conceptual and empirical readings of APT, and thereby accept that liberal political morality can in-principle and in practice accommodate a wider variety of restrictions on individual freedom than has been previously assumed.

Chapter 4

Stability and (Un)reasonability

The previous chapter established that the Absolute Prohibition Thesis (APT) prohibits policy means that fail to engage with persons as persons only when doing so is disrespectful to them. The permissibility of transformative manipulation therefore hinges on the existence of scenarios in which its use is justified on grounds of respect. My argument for their existence rests on the claim that liberal states have a respect-based duty to maintain the stability of political conditions under which rational persons are reasonably able to exercise basic autonomy. There are myriad ways that the stability of such conditions might be threatened. In the extreme, natural disasters, infectious disease, and foreign invasions can precipitate the collapse of public institutions. My focus, however, is on the threat that citizens themselves may pose to stability. Popper's paradox of tolerance looms large here: liberalism's basic political commitments constitute a doctrine of tolerance, but it cannot be a doctrine of *unlimited* tolerance. As he points out

> ...unlimited tolerance must lead to the disappearance of tolerance. If we extend unlimited tolerance even to those who are intolerant, if we are not prepared to defend a tolerant society against the onslaught of the intolerant, then the tolerant will be destroyed, and tolerance with them[1]

And yet liberal states cannot do whatever they want in the name of stability. They are still bound by the principle of respect for persons, and so cannot, for example, go around executing anyone who poses a threat. Respecting

[1] Popper 2013, 581n4. See Horton 1994 and Forst 2013, Ch.1 §1 for discussions of this and other paradoxes of toleration.

persons requires that they strike a balance between protecting the stability of political conditions under which rational persons are reasonably able to exercise basic autonomy, and minimizing the degree to which doing so interferes with persons' ability to exercise this capacity. However, whether the use of certain policy means, e.g. transformative manipulation, satisfies this requirement depends on the characteristics in virtue of which the relevant persons pose a threat to stability. Therefore, before we can evaluate whether there are any uses of transformative manipulation that defeat APT, we need a clear understanding of these characteristics and their connection to stability.

In §4.1 I analyse the concept of stability as the satisfaction of two criteria, compliance and enforcement, and how specifications of these criteria produce different models of what counts as stability for the *right reasons*. Using Rawls' account as an illustrative example, I argue that liberal models of stability fail to capture the different ways that individuals can contribute to or detract from the satisfaction of the compliance criterion. Consequently, they fail to fully specify the enforcement criterion - that is, whether and when the use of different stabilization mechanisms (i.e. policy means) such as transformative manipulation can be justified. In §4.2 I introduce a distinction between practical and doxastic (un)reasonability as a first step to addressing the gap in liberal models of stability. In effect, the distinction is between *conduct* that is(n't) consistent with respect for persons, and *normative commitments* that are(n't) not consistent with respect for persons. Finally, in §4.3 I combine these dimensions to produce four categories of persons - strongly reasonable, weakly reasonable, weakly unreasonable, and strongly unreasonable - and identify the kind of threat that persons falling into each category can pose to the stability of a liberal political order in virtue of these characteristics.

4.1 Political Stability

Stability is an important desideratum in the evaluation of political principles.[2] If we have good reason to doubt that a political order effectively governed by a set of principles will not reliably persist over time, then surely those principles must be rejected. In Rawls' view, which I adopt here, a political order is stable when, in the absence of external interference, it satisfies two basic criteria:

[2] Some deny this claim, e.g. Cohen 2008, pp. 327–328. I think these critics are mistaken, but leave the debate aside for my purposes here.

1. **Compliance**: individuals are willing to comply with the basic rules governing the political order, and more or less regularly do so

2. **Enforcement**: there exist stabilizing mechanisms to address infractions in a manner that promotes the satisfaction of (1)[3]

Note that these criteria can be satisfied in any number of ways. Individuals may comply with the rules of a political order for different reasons, e.g. fear, habit, endorsement, etc, and can come to have these reasons in different ways. Similarly, stabilization mechanisms can take as many forms as there are combinations of the features of policy means (instrument, method, mode, content), and their use can be restricted in various ways. As we'll see, only certain specifications of the two criteria - i.e. models - will be consistent with a given political morality. From this perspective, the model spells out when the associated political order is not *merely* stable, but stable for the right reasons.[4]

Does transformative manipulation have any place in a liberal respect-based account of stability for the right reasons? More specifically, can such policy means appear in the specified enforcement criterion? In this section I argue that no satisfactory answer can be given without articulating a clear picture of the ways that persons may pose a threat to the satisfaction of a suitably specified compliance criterion, and furthermore, that the most influential treatments of stability from a liberal perspective - most notably, Rawls' - fail to do so. I take Rawls' model of stability as my focus because it is by far the most developed within the liberal tradition, and provides the clearest illustration of the gap I wish to highlight. Furthermore, his account utilizes a concept - that of reasonability - that provides us with a fruitful way of framing the subsequent analysis of stability in this chapter.

4.1.1 Models of stability

Broadly speaking, models of stability fall into one of two categories: imposed and inherent.[5] Models of imposed stability condition the satisfaction of the compliance criterion on the satisfaction of the enforcement criterion. That is

[3] Rawls 1999, p. 6.

[4] I draw the distinction between stability and stability for the right reasons from Rawls 2005, pp. 458–462.

[5] I borrow this terminology from Weithman 2010, pp. 44–51. For reasons of simplicity my argument represents them as categorically distinct. As Klosko 2015 points out, the stability of actual political systems tends to involve some mixture of the two. This is not important for my purposes, however.

to say, individuals are willing to comply with the rules governing the political order only *because* effective stabilization mechanisms exist, but the existence and nature of those mechanisms does not depend on the willingness of individuals to comply with the rules. Hobbes provides us with a paradigmatic example of such a model: a political order is stable when individuals refrain from violating its basic rules because they regard reprisal by the sovereign as a credible threat.[6] If they cease to believe in the credibility of the threat - i.e. if they regard the sovereign as ineffectual - then they will (and indeed should) cease to adhere to the rules. To fail to do so would mean placing themselves under the protection of a sovereign that cannot protect, and therefore at the mercy of those who are not inclined to show them any.[7] Conversely, insofar as it preserves peace and order, the sovereign's ability to wield its authority in any manner it sees fit does not depend on the willingness of persons to abide by the rule governing the political order.

Clearly, imposed stability is inconsistent with the very spirit of the liberal tradition. From this perspective, the persistence of a political order should not depend on fear of reprisal for violating its rules - indeed, liberals have universally regarded this as a mark of illegitimacy. What they seek is a model of stability that inverts the relationship between the two criteria, i.e. the existence and nature of stabilization mechanisms is a product of individuals' willingness to more or less comply with the rules governing the political order. In other words, they seek a model of inherent stability. Rawls provides us with the most clearly articulated vision of stability in this sense.[8] Before discussing his specification of the compliance criterion, however, it is necessary to say something about the idea of reasonability. It is not only central to his model of stability, but will also be important for my forthcoming critique.[9]

There are two ideas of reasonability at play in the Rawlsian compliance criterion. The first is that of a reasonable comprehensive doctrine. A comprehensive doctrine is a set of what I have been referring to as normative

[6] Hobbes 2003, p. 96. Of course, Hobbes models the *creation* of the political order as a voluntary pact between individuals, but this does not bear on the question of the stability of that order once it comes into existence.

[7] By the First Law of Nature, individuals still have an obligation to seek peace, and by the Second defend themselves if this is not possible. The point is just that if the reigning political order does not secure this peace, which it cannot under an ineffectual sovereign, then they have no reason to abide by its terms.

[8] Rawls addresses the problem of stability throughout his writings. See Rawls 1999, Ch.8 and 2005, p. 140. See Barry 1995 for an excellent critical discussion of the evolution of Rawls' thought on the issue, as well as Weithman 2010, Ch.II.

[9] For a helpful discussion of Rawls' multifarious use of the expression, see Boettcher 2004.

commitments. In Rawls' words, these are broad frameworks of evaluation that address

> ...what is of value in human life, and ideals of personal character, as well as ideals of friendship and of familial and associational relationships, and much else that is to inform our conduct, and in the limit to our life as a whole[10]

It may seem implausible to assume that persons have a fully articulated schema of beliefs and attitudes on these subjects. For this reason, Rawls distinguishes between a *fully* comprehensive doctrine, which "...covers all recognized values and virtues within one rather precisely articulated system", and a *partially* comprehensive doctrine, which "...comprises a number of, but by no means all, non-political values and virtues and is rather loosely articulated".[11]

A comprehensive doctrine is reasonable when it satisfies three formal criteria: first, it is more or less internally consistent; second, it assigns weights to the values and beliefs it articulates and determines how conflict amongst them in their application is to be resolved; and third, though stable over time, it "...tends to evolve slowly in the light of what, from its point of view, it sees as good or sufficient reasons".[12] In sum, it is neither inconsistent, disordered, inflexible, or capricious. Note that the reasonability of a comprehensive doctrine does not hinge on its substantive content. Thus, even a morally repugnant set of normative commitments can be reasonable in the sense discussed here.

The second idea is that of a reasonable *person*. Such persons have two 'moral powers': the capacity for a conception of the good, and to form and revise a conception of justice.[13] More importantly for our purposes, however, they affirm a principle of reciprocity that prescribes a willingness to propose fair terms of cooperation - i.e. terms that are acceptable to others on the basis of their own commitments - and abide by them when others are willing to do likewise.[14] In addition, they also recognize the 'burdens of judgement', i.e. that for a variety of reasons, persons living under free institutions will inevitably come to affirm different reasonable comprehensive doctrines.[15] As

[10] Rawls 2005, p. 13. This is in contrast to a purely political doctrine, which pertains only to a society's primary political, economic, and social institutions ibid., p. 11.
[11] Ibid., p. 13.
[12] Ibid., p. 59.
[13] Ibid., p. 19.
[14] Ibid., pp. 49–50.
[15] Ibid., pp. 54–58.

such, they accept the (empirical) fact of reasonable value pluralism as something to be accommodated by a political order, rather than as a defect to be remedied through state action.[16].

With these ideas in hand, we can now return to Rawls' model of inherent stability. In his view, the satisfaction of the compliance criterion follows from agreement on just political principles amongst reasonable persons, each on the basis of their own reasonable comprehensive doctrine - that is, through a reasonable overlapping consensus.[17] For persons who are party to an overlapping consensus of this kind, compliance with the rules that govern the resulting political order constitutes an affirmation of their deepest commitments, rather than a compromise or necessary evil. Note that unlike the Hobbesian compliance criterion, the satisfaction of the Rawlsian compliance criterion makes no essential reference to the enforcement criterion. Rather, the satisfaction of the latter is a consequence of the former. Consider Rawls' claim that

> ...this feature of liberalism [i.e. how the compliance condition is satisfied] connects with the feature of political power in a constitutional regime: namely, that it is the power of equal citizens as a collective body[18]

Since the constitutional regime is an expression of the overlapping consensus, the latter is itself a stabilization mechanism. Put another way, effective institutional mechanisms for addressing threats to stability only exist in virtue of the overlapping consensus. Reasonable persons affirm the principles governing the regime for reasons that each identifies with, as so are willing to defend it against threats to its stability through their support for institutional mechanisms that serve this purpose.

4.1.2 Stabilization mechanisms

The relationship between the satisfaction of each criterion plays an important role in determining whether and when the use of different kinds of of stabilization mechanisms can be justified. In Hobbes' model of imposed stability, the sovereign is constrained only by the weakest of proportionality criteria:

[16] Rawls 2005, p. 60.

[17] On the idea of an overlapping consensus, see ibid., Lec. IV §3-7. Note that the overlapping consensus is a criterion of stability, not justice. For Rawls, the moral status of a set of political principles is determined by way of choice in the original position.

[18] Ibid., p. 144.

> It belongeth also to the Office of the Soveraign, to make a right application of Punishments...And seeing the end of punishing is not revenge...but correction, either of the offender, or of others by his example; the severest Punishments are to be inflicted for those Crimes, that are of most Danger to the Publique; such as those that proceed from malice to the Government established; those that spring from contempt of Justice[19]

The extreme latitude afforded to the sovereign here is a consequence of the primacy of the enforcement criterion. If its satisfaction is the reason *why* people comply with the rules governing the political order, then any restrictions on stabilization mechanisms cannot be derived *from* their (motivating) reasons for complying. All that matters is that peace and order are secured in an effective manner.

Models of inherent stability present us with a more complex picture. In Rawls' case, the question of whether and when the use of specific stabilization mechanisms can be justified depends on the target. While reasonable persons never intentionally seek to undermine a political order governed by principles that are derived from a reasonable overlapping consensus of which they are a part, they might nevertheless inadvertently threaten its stability. Perhaps their lifestyle choices indirectly erode the conditions for the very possibility of an overlapping consensus, e.g. economic prosperity, environmental sustainability, etc. In such cases, whether and when the use of different stabilization mechanisms can be justified is determined by the content of reasonability itself:

> [W]hen may citizens by their vote properly exercise their coercive political power over one another when fundamental questions are at stake?...only when it is exercised in accordance with a constitution the essentials of which all citizens may reasonably be expected to endorse in the light of principles and ideals acceptable to them as reasonable and rational[20]

In other words, coercion is a legitimate tool for addressing threats to stability, but only when it is consistent with just political principles that are the focus of an overlapping consensus amongst reasonable persons. Recall that amongst other attributes, reasonable persons (1) accept the fact of reasonable value pluralism, and (2) affirm the principle of reciprocity. This means that coercion cannot be used to maintain stability by suppressing or

[19] Hobbes 2003, p. 240.
[20] Rawls 2005, p. 217.

differentially favouring specific reasonable doctrines, for this would be inconsistent with (1). Nor can it involve imposing terms of cooperation that disadvantage certain groups of reasonable persons, for this would violate (2). Fortunately, neither of these measures would ever be necessary. Reasonable persons already endorse (1) and (2). If they threaten stability, then it is because they either misunderstand what their commitments entail, or they are unaware of the destabilizing impacts of their actions. Correcting these things only requires bringing them to recognize the inconsistency of their actions and commitments. If this cannot be accomplished solely through persuasion, then all else being equal, coercive measures are justified.

What about unreasonable persons who pose a threat to stability - that is, those who either reject the principle of reciprocity and/or the fact of reasonable value pluralism? Rawls' treatment of this question throughout his oeuvre is surprisingly thin. The most direct comment on the matter appears in a brief footnote in *Political Liberalism*:

> That there are doctrines that reject one or more democratic freedoms is itself a permanent fact of life, or seems so. This gives us the practical task of containing them - like war and disease - so that they do not overturn political justice[21]

His reasoning seems to be this: as noted, the only restriction on the state's use of stabilization mechanisms is consistency with the contents of an overlapping consensus amongst reasonable persons on just political principles. This means (a) no suppression of reasonable doctrines; and (b) no imposing disadvantages on persons who are willing to honour fair terms of cooperation. Note that neither (a) nor (b) places any restrictions on the suppression of *unreasonable* doctrines or persons. It follows that where these things have a sufficiently destabilizing influence, the state has license to contain them.

Rawls' specification of the enforcement criterion is incomplete in two respects. First, recall the different features and specifications of policy means from Chapter 1 (Table 4.1). As seen above, when applied to the case of reasonable persons he reduces the question of whether and when the use of different kinds of stabilization mechanisms is legitimate to the question of when the use of coercive stabilization mechanisms is legitimate. But this leaves unresolved whether, for example, the coercion must only be rational or if it can be utilized in a manipulative manner; whether it must take the target's normative commitments as given or if it can seek to affect a change to them; and the kinds of motivating reasons that the coercion is designed

[21] Rawls 2005, 64n19. See also Rawls ibid., p. 489.

Table 4.1: Features and specifications of policy means

Instrument	**Method**	**Mode**	**Content**
Persuasive	Rational	Conservative	Prudential
Coercive	Manipulative	Transformative	Moral
			Epistemic

to give rise to in the target. Things are even less clear when it comes to the treatment of unreasonable persons, for how are we to understand the idea of 'containment'? Unlike the Hobbesian sovereign, a liberal state does not have carte blanche in its use of stabilization mechanisms, for it is still bound by the principle of respect for persons, even unreasonable ones. In an effort to fill out this aspect of Rawls' account, Jonathon Quong defines containment as "any policy whose *primary intention* is to *undermine or restrict the spread of ideas* that reject [liberalism's] fundamental political values".[22] Such measures are legitimate, he argues, when applied to cases involving the inculcation of illiberal values, e.g. the spread of hate speech, the education of children.[23] While this furnishes us with a general definition of containment and two examples of legitimate use, it brings us no closer to answering the central question of whether and when the use of different kinds of stabilization mechanisms against unreasonable persons is legitimate.

These gaps in Rawls' specification of the enforcement criterion are related to the second source of incompleteness: the failure to explicitly acknowledge the different ways that individuals can pose a threat to the satisfaction of the compliance criterion. As we've seen, in his model of stability individuals are either reasonable or unreasonable. Reasonable persons can only inadvertently undermine the satisfaction of the Rawlsian compliance criterion since they are party to the overlapping consensus that defines it (i.e. if they knowingly sought to undermine it then by definition they would not be party to the overlapping consensus). Conversely, unreasonable persons may undermine the satisfaction of the compliance criterion precisely because they *reject* the Rawlsian specification. In other words, they reject the idea that stability for the right reasons can only be achieved through a reasonable overlapping consensus. But this picture glosses over the fact that people can evince different kinds of reasonability and unreasonability. It is at least conceivable that an individual may endorse a comprehensive doctrine that is incompatible

[22] Quong 2011, p. 299.
[23] Ibid., pp. 301–305.

with the definition of a reasonable person, and yet be willing to abide by the practical implications of the definition in the public sphere. Or, endorse a comprehensive doctrine that is consistent with the definition of a reasonable person, but reject its political implications.

Why is this distinction relevant? Because a complete specification of the enforcement criterion - that is, one that tells us whether and when the use of various combinations of the features of policy means in Table 4.1. constitute legitimate stabilization mechanisms - depends on a clear understanding of how different combinations of reasonability and unreasonability may undermine the satisfaction of the compliance criterion. The nature of the threat in these terms plays a crucial role in determining when the use of specific kinds of policy means - in particular, transformative manipulation - is consistent with the principle of respect for persons.

4.2 Conceptions of Reasonability

As we've seen, Rawls argues that a political order is stable for the right reasons only if its governing principles are affirmed by an overlapping consensus - that is, a joint affirmation of just political principles by reasonable persons, each on the basis of their respective reasonable comprehensive doctrine. Some commentators have argued that his model of stability begs the question. Brian Barry, for example, asserts that

> Rawls rigs the argument by saying that the condition of stability is that "the reasonable doctrines endorse the political conception, each from its own point of view" (PL, p. 134). But it is people, not doctrines, that go around endorsing conceptions...Rawls's way of putting it tacitly assumes the very point that is at issue: it presupposes that people can endorse principles of justice only if their "comprehensive view" endorses (in other words, entails or supports) them[24]

In other words, Rawls has not shown that inherent stability can *only* be secured through an overlapping consensus, for he has assumed, rather than demonstrated, that people cannot affirm just political principles that are at odds with their comprehensive doctrine. Barry's remarks highlight something of a tension between Rawls' concept of a reasonable person and his model of stability that is relevant to us here. Recall that such persons (a) affirm the principle of reciprocity and (b) accept the fact of reasonable value pluralism.[25]

[24] Barry 1995, p. 898.
[25] They also possess the two moral powers, but I leave this aside.

Suppose there are persons who, despite failing to jointly satisfy (a) and (b), are nevertheless willing to affirm political principles that are affirmed by anyone who does satisfy (a) and (b). If we accept Rawls' definitions, then the former are no less unreasonable than those who *reject* principles affirmed by the latter, and therefore pose the same kind of threat to the stability of the political order. But how can this be? Surely the fact that they affirm the principles governing this order means that any threat they pose to its stability is qualitatively distinct. If so, then are they covered by the same restrictions on the use of stabilization mechanisms against reasonable persons? Or are they still subject to 'containment'?

This tension in Rawls' model of stability suggests the need for a more fine-grained treatment of what it can mean to be a reasonable or unreasonable person. Kelly and McPherson, for example, draw a distinction between political and philosophical reasonability.[26] A person is politically reasonable when they satisfy (a) - that is, they are willing to propose fair terms of cooperation and abide by them insofar as others are willing to do likewise. Philosophical reasonability, on the other hand, means satisfying (b) - that is, recognizing that "...persons who engage in rational, critical reflection may come to disagree in deep and irreconcilable ways about the nature of the good".[27] George Klosko draws a similar distinction between attitudinal and cognitive reasonability.[28] The former is described in terms very much like political reasonability - attitudinally reasonable persons "...try to get along with others. They are open-minded and fair, not demanding more than their share...they are willing to live with others on fair terms of co-operation."[29] Echoing the idea of philosophical reasonability, a cognitively reasonable person is one whose "...beliefs or opinions are adequately grounded. She has good evidence...for holding them; her main principles are securely founded, while others, derived from them, follow according to sound rules of inference".[30]

What these distinctions get right is that they pick up on the contrast between a person evincing reasonable *conduct* versus reasonable *belief*. However, they draw the distinction in a way that fails to capture the full scope of what it can mean to be a reasonable or unreasonable person, and therefore the different ways that persons can pose a threat to stability. For example, philosophical unreasonability consists in endorsing a comprehensive doctrine

[26] Kelly and McPherson 2001.
[27] Ibid., p. 44.
[28] Klosko 2004.
[29] Ibid., p. 20.
[30] Ibid., p. 20.

that fails to recognize the fact of reasonable value pluralism. But one can fail to recognize this fact through one's actions, even if it is affirmed by one's comprehensive doctrine. Instead of carving up the concept of a reasonable person into a practical and a doxastic component, I'd like to treat each component as satisfiable in either practical *or* doxastic terms. In other words, I wish to draw the a more general distinction between practical and doxastic (un)reasonability.

For the purposes of the following discussion, it is helpful to connect Rawls' definition of a reasonable persons to the principle of respect for persons discussed in Chapter 3. Recall that this principle asserts that (i) persons have equal unconditional value *qua* persons; and (ii) this value is realized in part when rational persons are reasonably able to exercise basic autonomy.[31] Now, the principle of reciprocity can be understood as an expression of (i). The willingness to propose fair terms of cooperation and abide by them if others are willing to do likewise is a willingness to at the very least *treat* persons as having equal moral standing, with interests that are no less worthy of consideration than one's own. Accepting the fact of reasonable value pluralism, on the other hand, can be understood as an expression of (ii). Accepting that people living under free institutions will inevitably come to endorse different values means at least *treating* the (actual or potential) ability to exercise basic autonomy as one way in which the value of persons *qua* persons is realized. Recasting Rawls' definition of a reasonable person in terms of the principle of respect for persons allows us to draw more general conclusions about when persons pose a threat to the stability of a liberal political order, and not just a Rawlsian liberal political order.

4.2.1 Practical reasonability

Practical reasonability is a disposition to conduct oneself in the public domain in a manner that is consistent with the political principles that express respect for persons. A practically reasonable person need not endorse a comprehensive doctrine that actually affirms the unconditional value of persons or that this value is realised in part when rational persons are reasonably able to exercise basic autonomy. They need only conduct themselves as though they do affirm such a doctrine within the public sphere.

A person is practically unreasonable, then, when their conduct is incompatible with the principle of respect for persons. This can take a variety of forms. For example, one's conduct may be incompatible with the idea

[31] See §3.3

of persons having equal and unconditional value, while nevertheless being compatible with the claim that their value as persons is realized by being reasonably able to exercise basic autonomy. Will Kymlicka's discussion of the millet system in the Ottoman empire offers an interesting example of a political order that is practically unreasonable in this sense.[32] Following their conquests in the 14th and 15th centuries, the Ottomans came to rule over a large number of Christian and Jewish subjects. However, rather than attempting to eliminate non-Islamic creeds, they allowed these communities to not only practice their own religion, but also to manage their internal affairs in accordance with their own legal traditions.[33] Nevertheless, non-Muslims were forbidden from proselytizing, were subject to special taxes, had to wear distinctive dress, and faced restrictions on intermarriage. Thus, although the Ottoman's conducted themselves towards these communities in a manner that is consistent with valuing the ability to exercise basic autonomy, their willingness to impose the millet system on these communities is inconsistent with the idea of persons as having equal and unconditional value.

The second case of practical unreasonability involves a willingness to conduct oneself in a manner that is consistent with the idea of persons as having equal and unconditional value, but inconsistent with the ability to exercise basic autonomy as fundamental to realizing this value. We find a clear example in the writings of Locke. The pre-political state of nature is characterized as a state of perfect freedom, and of equality "...wherein all the Power and Jurisdiction is reciprocal, no one having more than another...".[34] It is through mutual consent between free and equal persons that political power is legitimized.[35] Since Locke is arguing against a background of religious pluralism, consent involves compromise, and therefore toleration, between persons who differ in their beliefs. However, Locke (in)famously excludes atheists from the scope of toleration, remarking that

> ...Those are not at all to be tolerated who deny the being of a God. Promises, Covenants, and Oaths, which are the Bonds of Humane Society, can have no hold upon an Atheist...those that by their Atheism undermine and destroy all Religion, can have no pretence of Religion whereupon to challenge the Privilege of a Toleration[36]

[32] Kymlicka 1996.
[33] Such self-governing communities were referred to as 'millets'.
[34] Locke 2013, II.§4.
[35] I leave aside the perennial debate about how to interpret Locke's claims about consent and their plausibility given his theoretical aims.
[36] Locke 2016, p. 159.

Despite appearances, this is not a denial of the equal and unconditional value of persons. Locke does think that we should not accept the presence of certain values and beliefs in the community, and indeed that those who endorse these beliefs and values should be punished. However, the purpose of this punishment is not retribution. Rather, Locke's view is that we must "...motivate the atheists to turn their eyes to their primary duty as men".[37] It is because they have unconditional value as persons that they are worth saving from values and beliefs that hinder the realization of this value. In a very real sense, Locke accepts certain restrictions on individuals' ability to exercise basic autonomy precisely *because* he regards persons as having equal and unconditional value.

4.2.2 Doxastic reasonability

Doxastic reasonability is the affirmation of a comprehensive doctrine that either entails the principle of respect for persons, or is at least consistent with it. Just as practical reasonability implies nothing about the content of comprehensive doctrines, so too is doxastic reasonability independent of one's conduct. It is a matter of beliefs and attitudes, not actions.[38]

Doxastic unreasonability has formal and substantive variants. The formal variants involve a failure to satisfy at least one of the criteria Rawls discusses in the context of a reasonable comprehensive doctrine. It may be unreasonable because it is inconsistent or unintelligible, e.g. certain fundamental duties cannot be satisfied without violating others, the duties it specifies are in-principle impossible to satisfy, or its criterion of truth or what counts as a good reason for belief is contradictory, etc. Or, it may fail to specify how conflicts between values are to be resolved, and is therefore insufficiently action-guiding. For example, in the absence of the difference principle, Rawls' conception of justice would render any comprehensive doctrine of which it is a part formally unreasonable because it would lack the resources to resolve conflicts between liberty and equality in the public domain.[39] Or, it could be such that, by its own lights, there is nothing that could justify modifications

[37] Numao 2013, p. 269. See also Waldron 2010, Ch.8.

[38] This should be taken as an analytical rather than psychological point. If someone consistently acted contrary to their professed beliefs we would be warranted in questioning their sincerity.

[39] Roughly speaking, the difference principle states that social and economic inequalities are justified only insofar as the associated benefits disproportionately accrue to the least well-off. See Rawls 1999, Ch.2 on the nature and role of the difference principle in his account.

to the doctrine no matter how minor. Religious fundamentalists typically endorse doctrines that display this feature. No part of their content is up for debate, or even *could* be up for debate. Indeed, to question or modify such doctrines would be tantamount to rejecting them altogether.

Substantive variants of doxastic unreasonability involve the affirmation of a comprehensive doctrine that is incompatible with respect for persons. Such a doctrine may regard persons as having value only in virtue of their achievements, or their service to a higher cause. An extreme example of the latter is expounded in *Kokutai no Hongi* (Cardinal Principles of the National Entity of Japan), a mytho-philosophical propaganda tract from 1937 that was widely disseminated in Imperial Japan.

> Loyalty means to reverence [*sic*] the Emperor as [our] pivot and to follow him implicitly. By implicit obedience is meant casting ourselves aside and serving the Emperor intently. To walk this Way of loyalty is the sole Way in which we subjects may "live", and the fountainhead of all energy. Hence, offering our lives for the sake of the Emperor does not mean so-called self-sacrifice, but the casting aside of our little selves to live under his august grace and the enhancing of the genuine life of the people of a State[40]

On this view, the individual has value only as a conduit of the Emperor's will. They have no interests that are not his interests, no duties that are not ultimately duties to him, and no claims to anything that he has not bestowed on them. This is not just a repudiation of persons as unconditionally valuable, but a denial of their very status as persons in any normatively salient sense.

Another substantive variant of doxastic unreasonability is the affirmation a comprehensive doctrine that denies that a person's being reasonably able to exercise basic autonomy is an essential part of how their value as a person is realized. In other words, it denies that the ability to exercise basic autonomy is constitutive of freedom. We see something of this in the political writings of Rousseau. There can be little doubt that his vision of a just society is motivated by a strong commitment to the inherent dignity of persons. The loss of natural freedom and equality following the emergence of *amour propre* is a tragedy for him because the relationships of dependence it engenders are an affront to what makes us distinctly human, i.e. our capacity for independent choice.[41] His attempt to recover freedom and equality within civil society given the challenge that *amour propre* poses is therefore nothing

[40] Hall 1949, p. 80.

[41] The centrality of independent choice in this respect comes out quite clearly in the following passages: "I can discover nothing in any animal but an ingenious machine, to

less than an attempt to discover the conditions under which the humanity of each and every person is most fully realized.[42] None of this would be intelligible if he did *not* view persons as unconditionally valuable.

Autonomy in the guise of moral freedom - understood as "...obedience to a self-prescribed law" - is central to Rousseau's solution to this problem.[43] What makes his a doxastically unreasonable doctrine, however, is that moral freedom requires the *exercise* of basic autonomy such that we commit ourselves to a kind of social virtue that is coterminous with the General Will. It is this feature of Rousseau's view that allows him to claim without contradiction that constraining persons to comply with the General Will is in effect forcing them to be free.[44] But this implies a principle of respect that is wholly incompatible with the one that liberals endorse, for the latter is premised on the claim that the unconditional value of persons is realized (in part) by their being *able* to exercise basic autonomy should they choose to do so. By this standard, forcing them to adopt or otherwise live in accordance with a narrow set of normative commitments infringes on basic autonomy, and is disrespectful to them as persons.

4.2.3 Summary

I began this section by arguing that Rawls' concept of a reasonable person does not provide a sufficiently nuanced picture of how persons may threaten the stability of a political order, and so does not support a complete specification of the enforcement criterion - that is, whether and when the use of different stabilization mechanisms is justified. To address this problem, I have proposed a distinction between practical and doxastic reasonability. In the next section, I examine the connection between a person's evincing some combination of practical and doxastic (un)reasonability and the threat they

which nature has given senses to wind itself up, and guard, to a certain degree, against everything that might destroy or disorder it. I perceive the very same things in the human machine, with this difference, that nature alone operates in all the operations of the beast, whereas man, as a free agent, has a share in his. One chooses by instinct; the other by an act of liberty...it is not therefore so much the understanding that constitutes, among animals, the specific distinction of man, as his quality of a free agent" (Rousseau 2002b, p. 95).

[42] We see this in Rousseau's stated aim in *The Social Contract* "To find a form of association that may defend and protect with the whole force of the community the person and property of every associate, and by means of which each, joining together with all, may nevertheless obey only himself, and remain as free as before" (ibid., p. 163).

[43] Ibid., p. 167.

[44] Ibid., p. 166.

may pose to the stability of a liberal political order, thus setting the stage in the final chapter for consideration of whether the use of transformative manipulation can be justified in any of these cases.

4.3 Patterns of (Un)reasonability

The forthcoming analysis of the threat that persons may pose to stability makes three assumptions. First, all persons evince some combination of practical and doxastic reasonability/unreasonability. This does not mean that they actually interpret their own actions and beliefs as being either consonant with or in opposition to respect for persons, only that they are capable of understanding it as such, e.g. if it were explained to them. There is a connection here to Rawls' claim that reasonable persons must possess the capacity for a conception of the good and for a conception of justice. Anyone who lacks these 'moral powers' is not capable of respecting or failing to respect others as persons, for this presumes a capacity to conceptualize moral value or worth.

Second, practical and doxastic (un)reasonability are more or less stable attributes of persons. People do not flit between patterns of conduct that are consistent with respect for persons and those that are not, nor do they easily change their beliefs and values in ways that fundamentally alter the consistency or inconsistency of their comprehensive doctrine with the principle of respect for persons. This is not to say that drastic changes never occur, only that they don't occur with great frequency. But neither is the expression of practical and doxastic (un)reasonability otherwise static. I adopt Rawls' assumption that a person's comprehensive doctrine "...tends to evolve slowly in the light of what, from its point of view, it sees as good or sufficient reasons"[45], as do a person's reasons for conducting themselves in accordance with or in opposition to the principle of respect.

Third, there is no necessary connection between practical and doxastic (un)reasonability. As we saw in the previous section, Rawls' model of stability explicitly assumes that people will only affirm political principles that are supported by their comprehensive doctrine. This amounts to the claim that practical reasonability entails doxastic reasonability. I make no such assumption here. As we'll see, there is nothing psychologically incoherent about cases where the two dimensions of (un)reasonability clash. Furthermore, these cases arguably make up the largest classes of potential threats

[45] Rawls 2005, p. 59.

to stability. I return to this point below.

Putting the two dimensions of (un)reasonability together produces four categories of persons that I refer to *strongly reasonable*, *weakly reasonable*, *weakly unreasonable*, and *strongly unreasonable*.[46] In the following sections I consider each of these categories in turn with the aim of identifying the kind of threat that their members may pose to the stability of a liberal political order.

4.3.1 Strong reasonability

Strongly reasonable persons evince both practical and doxastic reasonability. Their conduct in the public domain affirms liberalism's basic political commitments, and they endorse a doctrine that is at least consistent with the idea of persons as unconditionally valuable. There are two ways that this description may be satisfied. In the first case, an individual endorses liberal political commitments as the political expression of a comprehensive doctrine that explicitly affirms the principle of respect. For example, one might affirm the treatment of persons as free and equal in the political domain because, like Kant, one regards persons as having an inherent dignity that places weighty restrictions on how they can be treated. In the second case, an individual endorses liberal political commitments because they create conditions that are conducive to realizing values that, although consistent with respect for persons, do not explicitly endorse the principle. For example, one might endorse a religious doctrine according to which persons are created as free and equal in the eyes of God, and cannot find salvation except through the exercise of their capacities as rational creatures. This doctrine does not include the principle of respect for persons *qua* persons. Rather, respecting persons is a means of respecting their creator. In practice, this means ensuring that people have the ability to fulfil what god desires for them, namely, to find salvation through the free exercise of their rational capacities. They affirm liberal political commitments because they create the conditions that make this possible. What unites these cases is that practical reasonability is motivated for reasons that are internal to a person's comprehensive doctrine.[47]

Strongly reasonable persons are unlikely to pose a threat to stability be-

[46] This taxonomy draws to some degree on Klosko's 2004 distinction between strong and weak reasonability.

[47] As these examples indicate, the agents who appear in Rawls' overlapping consensus are strongly reasonable persons, and are therefore central to his account of stability for the right reasons.

cause their practical restraint in the political realm is motivated by a deep commitment to comprehensive values that are at least consistent with respect for persons. However, there are at least two scenarios where they may have a potentially destabilizing influence. First, they may believe that their actions are consistent with support for liberal political conditions, when in fact they are not. For example, someone might not understand that the economic policies of the candidate that they intend to vote for will create disparities that are incompatible with a regime of equal rights. Or, they may not recognize that their lifestyle choices contribute to conditions that undermine the feasibility of a liberal political order. A pressing concern of this kind is the aggregative impact of individual consumption choices on climate change.

Second, strongly reasonable persons may be politically apathetic. They endorse a liberal political order for reasons internal to their comprehensive doctrine, but are happy to live in ignorance of its operations. Perhaps they fail to exercise their vote during elections, pay little attention to consequential policy changes, or simply make no effort to understand how its most important institutions function. The worry here is that without knowledge and active support from citizens, a liberal political order is particularly vulnerable to degradation from within through, for example, corruption or capture by special interests.[48]

4.3.2 Weak reasonability

Weakly reasonable persons endorse comprehensive doctrines that are consistent with the principle of respect, and yet affirm political principles that conflict with it. In other words, they affirm that persons have unconditional value that is realized in part by their being reasonably able to exercise basic autonomy, but favour political conditions that are inconsistent with one or both of these things. For example, they might argue that political equality can only lead to kakistocracy, a worry that Mill highlights in his remark that

> The natural tendency of representative government...is towards collective mediocrity: and this tendency is increased by all reductions and extensions of the franchise, their effect being to place the principal power in the hands of classes more and more below the highest level of instruction in the community[49]

[48] This is of course an animating concern of republicans and republican-minded liberals. For a useful discussion of the problem of apathy, see Dagger 1997, pp. 133–135.

[49] Mill 2015, p. 273.

If one were to think this problem inescapable (which Mill does not) and that it could only lead to conditions that do not afford persons the ability to exercise basic autonomy, then one might reject liberal political conditions. We see hints of weak reasonability in the 'enlightened despots' of 18th and 19th century Europe. Although no fan of political or social equality, Frederick the Great defines the sovereign's duty in terms of their subjects interest in something that appears to be very close to basic autonomy:[50]

> ...if we return to the early origins of society, it would appear that the sovereign has no right to determine the thinking of his subjects. One would have to be insane to imagine that men have ever said to a fellow man, 'We are raising you above ourselves, because we like to be slaves, and we are giving you the power to direct, as you will, our thoughts.' On the contrary, they said, 'We need you to uphold the laws we wish to obey, to govern us wisely, and to defend us; we demand of you, moreover, that you respect our freedom.'[51]

Without too much distortion, weak reasonability can be understood as a kind of paternalism. The weakly reasonable person affirms the unconditional value of persons and the importance of basic autonomy for the realization of this value, but does not trust their ability to establish and maintain political conditions that make the realization of this value possible.

The very fact that weakly reasonable persons reject liberal political commitments makes them a potential threat to stability. They would prefer an illiberal order that, in their view, would do a better job of creating conditions under which persons are reasonably able to exercise basic autonomy - say, some form of technocracy or epistocracy.[52] But any efforts in this direction on their part are constrained by their endorsement of a comprehensive doctrine that includes the principle of respect. They do not attempt to bring about change through coercion, for example. It would be characteristic of such persons to work within the political system to bring about the change they seek through consent.

Against the backdrop of a stable liberal order, however, their efforts are likely to meet with failure, which brings us to a potentially greater threat that weakly reasonable persons can pose to stability. It is not implausible that the frustration of repeated failures over time can lead to the development of personal characteristics that social psychologists have linked to radicalization,

[50] I am of course leaving aside the question of whether Frederick the Great ever actually lived up to the ideal he espoused.

[51] Frederick 2021, p. 205.

[52] See Brennan 2016 for a defence of epistocracy.

e.g. feelings of victimization and political grievance; self-persuasion into increasingly radical ideas and acts; and increasing animus towards opponents.[53] Political radicalization, in turn, is linked with a tendency to dehumanize the other, and therefore cease to see them as worthy of respect as persons. No longer constrained by a comprehensive doctrine that is at least consistent with respect for persons, they are more likely to pursue their political aims through means that are antithetical to the principle, e.g. threats, violence, etc. The weakly reasonable person is therefore vulnerable to tipping over into strong unreasonability, which, as we'll see below, presents the greatest potential threat to stability.

4.3.3 Weak unreasonability

Weakly unreasonable persons endorse a comprehensive doctrine that denies the principle of respect for persons, but they are nevertheless willing to support a liberal political order. In other words, they support a liberal political order, but *not* for reasons that are internal to their comprehensive doctrine. For example, they may do so because imposing their doctrine on the wider polity isn't a realistic goal, and they have no wish to have others' doctrines imposed on them. Theirs is, to borrow an expression from Judith Shklar, a liberalism of fear.[54] Or, they may believe that political cooperation with those who do not share their comprehensive doctrine is necessary to secure certain collective goods given the infeasibility of imposing their view on others.[55] Or they may respect the fact of a liberal political order because their comprehensive doctrine, which denies the principle of respect, is apolitical.

Such persons pose a potential threat to stability in virtue of the contingency of their support for a liberal political order. If it is motivated by the belief that imposing their doctrine on the wider polity is infeasible, then a change in circumstance may lead them to abandon their practical restraint. They "...accept the political conception as a mere *modus vivendi*...biding their time until the balance of power shifts in their favor, whether through sheer force or a tyranny of the majority".[56] A similar worry applies in the case of those who are motivated by an apolitical doctrine that rejects the principle of respect. Such doctrines tend to be affirmed by minority religious communities that are strongly averse to pluralism within their ranks.

[53] Kruglanski, Bélanger, and Gunaratna 2019, p. 73.
[54] Shklar 1989.
[55] McCabe 2010 makes the case for a *modus vivendi* liberalism that is derived from something like this picture.
[56] Kelly and McPherson 2001, p. 54.

Their practical willingness to accept liberal political conditions is typically motivated by the protection it affords them from external persecution. However, whenever a community of such persons becomes sufficiently influential in the wider polity, it is liable to seek to impose its doctrine on others - in other words, apolitical doctrines tend to become politicized as they grow in influence.[57]

In summary, weakly ureasonable persons pose a potential threat to stability in virtue of the pragmatic nature of their support for liberalism's basic political commitments. In effect, the only thing separating them from strong unreasonability is a belief in the expediency, rather than morality, of compliance with liberal political commitments.

4.3.4 Strong unreasonability

Strong unreasonability denotes a self-conscious hostility to liberal political commitments in both conduct and belief. It is a comprehensive rejection of the principle of respect for persons. Recall that this principle prescribes, first, adopting attitudes/beliefs that acknowledge that the unconditional value of persons *qua* persons is realized in part by their being reasonably able to exercise basic autonomy; and second, conducting oneself in a manner that is consistent with the preservation of conditions under which individuals are reasonably able to do so. Strongly unreasonable persons reject the claim that persons are deserving of respect *qua* persons, and/or that their ability to exercise basic autonomy is at all relevant to what respect for persons requires. Furthermore, they are actively hostile to liberal political conditions regardless of the benefits that may accrue under them. Such persons might, for example, vote for candidates that pledge to roll back democratic or civil rights; or leverage civil associations such as think tanks or institutions of higher education to promote the spread of comprehensive doctrines that recognize no distinction between the political and the private; or use violence to further an ethno-nationalist cause. Katherine Stewart's description of Christian Nationalism in the United States provides a compelling portrait of strong unreasonability:

> Christian nationalism...asserts that legitimate government rests not on the consent of the governed but on adherence to the doctrines of a specific religious, ethnic, and cultural heritage. It demands that our

[57] As an illustration, see Cromartie 1996 for a wide-ranging discussion of changing interpretations amongst Christians of Jesus' injunction to "Render to Caesar the things that are Caesar's, and unto God the things that are God's".

> laws be based not on the reasoned deliberation of our democratic institutions but on particular, idiosyncratic interpretations of the Bible...It looks forward to a future in which its versions of the Christian religion and its adherents, along with their political allies, enjoy positions of exceptional privilege and power in government and in law[58]

Here we see a doctrine that rejects the view that individuals have, in virtue of their status as persons, equal basic rights that place very strong constraints on state action, and is actively hostile to any political order that espouses such a view. Of course, strong unreasonability comes in secular flavours as well. Consider the following description by Benito Mussolini and Giovanni Gentile (though almost certainly written entirely by the latter) of the relation between state and individual from the perspective of Fascism:

> The keystone of the Fascist doctrine is its conception of the State, of its essence, its functions, and its aims. For Fascism the State is absolute, individuals and groups relative. Individuals and groups are admissible in so far as they come within the State. Instead of directing the game and guiding the material and moral progress of the community, the liberal State restricts its activities to recording results. The Fascist State is wide awake and has a will of its own.[59]

As these passages illustrate, strongly unreasonable persons pose the greatest potential threat to stability in virtue of the uncompromising nature of their hostility to liberal political commitments as an expression of respect for persons. Unlike the weakly unreasonable person, they are not prepared to even provisionally accept a liberal political order as legitimate. And unlike the weakly reasonable person, their hostility to the political order is motivated by a self-conscious rejection of the view that persons have unconditional value that is realized in part by their being reasonably able to exercise basic autonomy.

However, there is another more indirect channel by which they may have a destabilizing influence. Their hostility to the political order and its underlying values has the potential to provoke hostility from others that transforms into strong unreasonability. For example, suppose that person A is willing to comply with the liberal political order despite rejecting its underlying values because they want protection from being forced by others to live in accordance with a comprehensive doctrine they do not endorse. In other words,

[58] Stewart 2020, Introduction, para 11.

[59] Mussolini and Gentile 1932. It should be noted that the authors describe Fascism as a religious doctrine, but it is obvious from their discussion that this is simply rhetoric.

A is weakly reasonable. B, on the other hand, is unwilling to accept the liberal political order *because* they think it fails to adequately realize the values that ostensibly underpin it. However, they do not seek to overthrow the order, but rather transition to another by rationally convincing others that they are right. B is therefore weakly unreasonable. Finally, C is not only unwilling to accept the liberal political order as legitimate, but actively seeks to replace it with one governed by a comprehensive doctrine that rejects the principle of respect, and through means that violate the principle if necessary. C is therefore strongly unreasonable. The worry here is that if A comes to regard C's presence within the political community as intolerable, they may advocate for policy measures that are incompatible with respect for persons, e.g. arresting C for purely political speech. Indeed, such measures may be a truer expression of A's comprehensive doctrine that the reigning liberal political order. Thus, C's influence could provoke A to abandon practical reasonability, thereby transitioning to strong unreasonability. Similarly, B could come to regard C's presence in the political community as intolerable, not because the latter is hostile to the political order *per se*, but because their motives are incompatible with respect for persons. We can imagine that B might paradoxically cease to regard C as worthy of respect as a person, thereby abandoning doxastic reasonability and becoming strongly unreasonable. Admittedly, this is conjecture, but not implausible conjecture. Reciprocity is built into the very fabric of morality for reasons that are as much psychological as philosophical.

4.3.5 Summary

In this section we've looked at four categories of (un)reasonability and the kinds of threats that persons falling into these categories may pose to the stability of a liberal political order. The purpose of this exercise has been to fill a gap in existing liberal models of stability when it comes to the specification of the enforcement criterion. Without a clear understanding of the different ways that persons can pose a threat to stability, it is impossible to evaluate whether and when the use of different stabilization mechanisms - and in particular, transformative manipulation - is justified. Having clarified the different categories of potential threats, we can now move on to the question of justification.

4.4 Conclusion

I began this chapter by noting that APT is false only if there are plausible scenarios in which, despite failing to engage with persons *as* persons, the use of transformative manipulation does not violate the principle of respect. A core part of my overall thesis is the claim that the state is justified in utilizing such policy means if it is necessary for maintaining the stability of a political order under which persons are reasonably able to exercise basic autonomy. As noted in the introduction, evaluating whether it *is* ever necessary requires that we have a clear picture of the kinds of characteristics that make persons a potential threat to stability. My aim in this chapter has been to produce such a picture.

I have defined stability in terms of the joint satisfaction of compliance and enforcement criteria, and stability for the right reasons in terms of a particular specification of these criteria (i.e. a model of stability). Liberals are committed to a model of inherent stability, wherein the satisfaction of the enforcement criterion is conditioned on the satisfaction of the compliance criterion. Using Rawls' account as an illustrative example, I have argued that liberal theorists have not provided a sufficiently nuanced account of how its satisfaction may be threatened, and so fail to provide a complete specification of the enforcement criterion - that is, whether and when the use of different kinds of stabilization mechanisms is justified.

Drawing on Rawls' concept of a reasonable person, I have modelled the potential threat that persons may pose to stability in terms of practical and doxastic (un)reasonability. In combination, these characteristics produce four categories of persons - strongly reasonable, weakly reasonable, weakly unreasonable, and strongly unreasonable - each of which is associated with a different kind of threat to stability. With this taxonomy in hand, we can now move on to evaluating whether the use of transformative manipulation to mitigate the destabilizing influence associated with each category of person can ever be justified.

Chapter 5

In Defence of Transformative Manipulation

Before outlining my aims in this final chapter it will be helpful to summarize how we have arrived at this point. Chapter 1 defined policy means as state actions that aim to bring persons to regard themselves as having sufficient reason(s) to act in ways that contribute to the achievement of the specific policy ends. How they are brought to do so depends on the specification of the policy means in terms of four characteristics: instrument, method, mode, and content. I argued that theorists have neglected the political morality of policy means in the transformative mode, and this despite these means having certain practical advantages. I then motivated my focus on transformative policy means that employ manipulation by noting that if it can be justified, then it is likely that less objectionable kinds of transformative policy means (e.g. transformative rational persuasion) can as well.

Chapter 2 introduced the Absolute Prohibition Thesis (APT) which state that no plausible conception of liberal political morality can accommodate the use of transformative manipulation. At its heart is the Fundamental Liberal Principle (FLP), which is the presumption that interference with individual freedom is morally wrong unless the interfering party can provide adequate justification. Since liberals endorse a variety of conceptions of freedom, FLP entails APT iff transformative manipulation conflicts with any plausible conception of liberal freedom, and in a manner that cannot overcome the presumption of non-interference on any interpretation. Through a descriptive analysis of the concept of freedom, I argue that this is equivalent to the claim that for any plausible conception of liberal political morality, transformative manipulation undermines the satisfaction of conditions that realize the value of agency in a manner that is morally impermissible.

Chapter 3 identified what I take to be the strongest argument for APT based on FLP so construed. For any plausible conception of liberal political morality, the ability to exercise basic autonomy is a constitutive part of realizing the value of agency - i.e. freedom. Transformative manipulation undermines the ability to exercise basic autonomy, and therefore interferes with any plausible conception of liberal freedom. The strongest argument for its impermissibility on these grounds is that it undermines basic autonomy in a manner that violates a principle of respect for persons *qua* persons that underpins liberalism's basic political commitments. I note, however, that if the state has a respect-based duty to protect the stability of political conditions under which persons are reasonably able to exercise basic autonomy, and discharging this duty sometimes requires treating people in ways that fail to engage with them as persons, then respecting persons doesn't *necessarily* require engaging with them as persons. This argument appears to refute APT as a conceptual claim. However, unless we can identify plausible cases where such treatment is necessary, then APT survives as a *de facto* thesis.

Chapter 4 analysed stability in terms of compliance and enforcement criteria, and highlighted how liberal models of stability condition the satisfaction of the latter on the satisfaction of the former. A complete specification of the enforcement criterion determines whether and when the use of different kinds of policy means to address different kinds of threats to stability can be justified. However, extant specifications of the compliance criterion do not provide a sufficiently nuanced picture of the characteristics in virtue of which persons may pose such a threat. Drawing on the twin concepts of practical and doxastic (un)reasonability, I have proposed a taxonomy of persons as either strongly reasonable, weakly reasonable, weakly unreasonable, or strongly unreasonable, and argued that each category is associated with a unique threat profile. The question, then, is whether the use of transformative manipulation to address any of these categories of threats can ever be justified on grounds of respect for persons.

In the present chapter, I defend an affirmative answer to this question. Doing so will require articulating precisely *when* persons in each category pose an actual (and not merely potential) threat to stability, then demonstrating that in at least some of these cases, the use of transformative manipulation to defuse the threat is consistent with respect for persons. This latter task, which comprises the bulk of my discussion, is guided by two constraints on the justification of transformative manipulation in the name of stability. First, its use is consistent with respect for persons only if alternative means which engage with persons as persons would not adequately address the threat. In other words, the use of transformative manipulation can only be justified as a last resort. And second, the substance and extent

of the modification to the target's normative commitments cannot exceed what is necessary to address the threat they pose to stability. The idea here is that it cannot be the case that once the first constraint is satisfied, the state has free reign to shape the target's normative commitments however it likes. The change must be tightly tied to the purpose at hand, namely, protecting the stability of a liberal political order.

In §5.1 I argue that the use of transformative manipulation against strongly unreasonable, weakly unreasonable, and weakly reasonable persons who pose a genuine threat to stability can be justified. For each case, I clarify when they pose such a threat, and argue that there will be at least some cases where addressing this threat plausibly requires the use of transformative manipulation. In the case of strongly unreasonable persons, I argue that transformative measures are justified in virtue of the intractability of the target's illiberal convictions, whereas manipulative methods are justified on the grounds that rational methods are liable to be self-defeating. In the case of weakly unreasonable persons, I argue that transformative measures are justified because, in light of their doxastic unreasonability, we cannot appeal to their existing normative commitments when they are prepared to abandon practical reasonability. Manipulative methods are justified on the grounds that their awareness of efforts to impose different normative commitments on them is likely to inflame, rather than mitigate, their willingness to abandon practical reasonability. And finally, in the case of weakly reasonable persons I argue that transformative measures are justified when their practical unreasonability is a product of the satisfaction of faulty epistemic commitments. Manipulative methods are justified because the target's awareness of attempts to affect a transformation to their epistemic commitments risks being perceived in a manner that solidifies their practical unreasonability.

In §5.2 I argue that in each of the three cases, any justified use of transformative manipulation must aim at strong reasonability. In doing so, I note an apparent tension with the requirement that any use of transformative manipulation is constrained by the requirement that the magnitude of the affected change is no greater than is necessary to address the relevant threat to stability. I argue that this is only an apparent tension which dissolves when we consider whether alternatives to aiming at strong reasonability are likely to effectively address the relevant threat to stability. Since we have compelling reasons to believe they do not, any aim but strong reasonability constitutes a failure of recognition respect, and so is impermissible.

Finally, in §5.3 I explain why these arguments, if sound, entail the abandonment of even the *de facto* interpretation of APT.

5.1 Three Cases of Justified Transformative Manipulation

As we saw in Chapter 4, each of the four categories of (un)reasonable persons poses a different kind of potential threat to the stability of a liberal political order. In this section, I aim to show that the use of transformative manipulation can be justified in three of these cases, namely, strong unreasonability, weak unreasonability, and weak reasonability. First, however, the omission of strong reasonability from this list requires some explanation.

Recall that strongly reasonable persons comply with the rules governing a liberal political order, and for reasons that are internal to a comprehensive doctrine that is at least consistent with the principle of respect for persons. If such persons pose a threat to the stability of a liberal political order, then it can only be inadvertently. As noted previously, this may because they fail to recognize how their choices contribute to social or environmental phenomena that undermine the sustainability of liberal political conditions. For example, ecological collapse due to climate change will almost certainly lead to widespread social and economic turmoil, the successful management of which is unlikely to accommodate a regime of individual rights that place robust restrictions on state action. The strongly reasonable person will naturally abhor this state of affairs not simply for prudential reasons, but also because it means the loss of conditions that reflect their most fundamental commitments. As such, it is not necessary to affect a change to their normative commitments to bring them to see themselves as having sufficient reason to modify their behaviour so as to avoid contributing to the problem. Certainly, it may be challenging to get them to understand or accept the impact of their choices, particularly when the effects are not immediately apparent. Solving this problem may require the use of manipulative rather than rational methods, and/or instruments of coercion rather than persuasion. But the purpose of these efforts is help them recognize the incongruity between their existing commitments and the impact of their choices. In other words, it requires effectively *leveraging* their existing commitments rather than affecting a change to them. Thus, insofar as we are dealing with strongly reasonable persons, there is no respect-based justification for anything but policies in the conservative mode. They never intentionally *seek* to destabilize the political order, and if their choices inadvertently do so, then it is sufficient to bring them to recognise what already matters to them. I now turn to why the same cannot always be said of the other three kinds of cases.

5.1.1 Strongly unreasonable persons

Recall that strongly unreasonable persons evince a hostility to liberal political conditions that is motivated by normative commitments that reject the principle of respect. Typical examples include religious zealots and political extremists of various kinds. When do such persons pose a genuine threat to stability? One might assert that they must exist in reasonably large numbers relative to the general population, whereas individuals or small groups are akin to nothing more than grains of sand in the workings of a great machine.[1] It follows that any justification for the use of transformative manipulation on strongly unreasonable persons is subject to a threshold condition below which the use of such policy means is impermissible.

This view should be rejected, for there is no straightforward relationship between the size of a population of strongly unreasonable persons relative to the general population and the magnitude of the threat they pose to stability. For example, small groups of fanatics may engage in terroristic acts that, although posing no material threat to stability in themselves, are designed to provoke a response that does exert a destabilizing influence, e.g. the curtailing of individual freedoms, inflamed intergroup tensions, etc. Or, small groups of strongly unreasonable persons may be willing and able to engage in acts that cause enormous destruction and/or loss of life. Technologies that make this possible proliferated during the Cold War - for example, the infamous 'suitcase nuke'[2] - and continue to do so. Additionally, individuals or small groups of strongly unreasonable persons can also threaten stability in a top-down manner. With sufficient resources, they may be able to influence legislation or affect appointments to key public offices in ways that erode the foundations of a liberal political order. For example, they might motivate legislators to pass a bill that disenfranchises certain classes of voters, or to appoint sympathetic individuals to judicial positions in order to bypass checks on unconstitutional policies.

Because strongly unreasonable persons need not be numerous to exert a destabilizing influence, the state is warranted in treating their mere presence as a threat to stability. Certainly, there may be individuals and groups who are not presently engaged in destabilizing projects, or who's projects are ineffective. But to say that for these reasons they do not pose a threat is to miss the forest for the trees. Their hostility to liberal political conditions creates a vulnerability that, if it is not already being exploited, could very easily be exploited at any time, e.g. by the emergence of a figure or movement

[1] See Quong 2011, pp. 303–304 for a discussion of this claim.
[2] See James Martin Center for Nonproliferation Studies 2002

that seeks to overthrow the liberal political order. Furthermore, as discussed in §4.3.4, their presence may also serve to radicalise weakly reasonable and weakly unreasonable persons. The claim that strongly unreasonable persons should always be treated as a threat to stability may sound rather extreme, even illiberal. It is not, however, insofar as the way they are treated is consistent with respect for persons, understood as engaging with them in a manner that is consistent with the ongoing realization of their unconditional value as persons. I now turn to what satisfying this constraint requires.

Why transformation?

Having established that strongly unreasonable persons are a threat to stability as such, we are now faced with the question of whether the use of transformative manipulation to mitigate this threat is ever justified. As noted above, this can only be the case if we have reason to believe that policy means that engage with persons as persons would not adequately address the threat. Two options present themselves here. The first is to utilize policy means in the conservative rather than transformative mode. The second is to apply rational rather than manipulative methods. I examine the former option here, before considering the latter in the next section.

Recall that policies in the conservative mode attempt to bring the target to regard themselves as having sufficient reason(s) to ϕ by appealing to their existing normative commitments. In the case of strongly unreasonable persons, ϕ represents something like 'complying with the rules governing a liberal political order'. What reason do we have for doubting the effectiveness of this approach? Most obviously, the fact that their conduct is already motivated by a comprehensive doctrine that is fundamentally hostile to liberalism's basic political commitments. Their practical unreasonability is a reflection of a profound doxastic unreasonability. Attempting to motivate practical reasonability by leveraging their existing normative commitments would, at best, only result in a kind of temporary practical incoherence.

To make things a bit more concrete, imagine that Bob believes that individuals have no value except as members of a certain political movement. They warrant respect to the degree that they are prepared to further its aims. Those outside of the movement, and those who betray its ideals from within, have no moral standing whatsoever. They may be used and discarded as necessary, for the weight of their individual interests is nothing relative to the interests of those who accept the movement's ideals and work to bring it to fruition. Bob lives in a liberal society that endows persons with all the familiar rights and freedoms. His commitments are antithetical to this political order, and for this reason is intolerable to him. He therefore goes about

attempting to convert others to the cause, supports acts of political violence, and so on. Now, it is plausible that Bob could be brought to regard his doctrine as giving him sufficient reason to *comply* with the existing political order? This seems contradictory. If sincere belief in the values encoded in the doctrine is fundamentally incompatible with acknowledging the legitimacy of a liberal state, then bringing about practical reasonability must necessarily involve a change of some kind his normative commitments.[3]

One might respond to this argument along the following lines: the liberal tradition finds its historical roots in an ideal of tolerance amongst adherents of rival political and theological doctrines.[4] Part of the explanation for its appearance is as a response to decades of ruinous conflict fuelled at least partly by disagreement on fundamental values. The intractability and costs of these conflicts created the impetus to seek political principles that accommodate a variety of views on matters of faith and conceptions of the good, if only for reasons of self-interest. The lesson is this: even if strongly unreasonable persons cannot be motivated to comply with the rules of a liberal political order by reasons internal to their comprehensive doctrine, they may at least be moved to do so by prudential reasons. Bringing them to recognize these reasons can be accomplished by policies in the conservative mode, since what they regard as their self-interest is determined in part by their normative commitments. The state therefore has an avenue for addressing the threat that strongly unreasonable persons pose to stability in a manner that engages with them as persons.

It is undoubtedly true that people can be practically reasonable for purely prudential reasons - indeed, this is what makes someone weakly unreasonable, a category I return to shortly. But this is precisely why the objection misses the mark. Our aim here is to determine if we have reason to believe that policies in the conservative mode might effectively address the threat posed by strongly unreasonable persons to stability. By hypothesis, their doxastic unreasonability is such that compromise with liberal political commitments is intolerable. It is the defining characteristic of the fanatic that they are not amenable to reasoned compromise, and certainly not for reasons of self-interest. Therefore, it won't do to claim that the state can always elicit the necessary prudential motivation by leveraging their existing commitments. There will inevitably be varieties of doxastic unreasonability for which reasons of prudence are as irrelevant as reasons of respect. If our aim is practical reasonability, the use of transformative measures cannot be avoided in such

[3] I examine the nature of this change in the next section.

[4] I am greatly simplifying the historical picture, but further nuance isn't important here.

cases.

Manipulation and (self-)respect

Transformative policy means do not necessarily fail to treat persons as persons. Certainly, there is an undeniable tension between targeting an individual's normative commitments for modification and recognizing them as a more or less rational agent who is capable of coming to their own conclusions about what to believe and how best to live their life. But we can still engage with them as persons if we pursue the desired change in a manner that treats them as the final arbiter on the matter. In other words, by employing rational rather than manipulative methods. Recall from Chapter 1 that policy methods are manipulative when the application of the relevant instruments (persuasive or coercive) and/or its rationale is intentionally obscured from the target.[5] Conversely, the method is rational if policymakers intend for the target to be aware of the application of the instrument and the purpose it is meant to serve. If stability-related uses of transformative manipulation against strongly unreasonable persons can be justified, then we must have reason to believe that rational methods are unsuitable.

One reason we might doubt the efficacy of rational methods in this case is that it is liable to provoke a backlash from the targets. Let us return to the example of Bob the political extremist. Bob's normative commitments are incompatible with the claim that persons have unconditional value *qua* persons; rather, any value they have must be earned through their commitment to the (illiberal) political cause. In other words, it follows that persons can only ever warrant appraisal respect.[6] As noted in the previous section, defusing the threat that Bob poses to stability requires affecting a change to his normative commitments that renders them at least consistent with the idea that persons have unconditional value *qua* persons, and that this value is realized in part by their being reasonably able to exercise basic autonomy. Therefore, they warrant recognition respect, which requires Bob to conduct himself in ways that are consistent with the preservation of their ability to do so.

Now, if the state utilizes rational methods - e.g. transformative rational persuasion, transformative rational coercion - then their efforts to bring about the requisite change to Bob's normative commitments will be transparent to him. In other words, he will be aware of the application of the policy instrument *as* a policy instrument, and he will understand the aims of the

[5] See §1.1.2.
[6] See §3.2.4 on the distinction between appraisal and recognition respect.

policy of which it is the means. It is not implausible that such efforts will be counterproductive, partly for reasons highlighted by Mill in his critique of paternalistic infringements on individual liberty:

> If there be among those who it is attempted to coerce into prudence or temperance, any of the material of which vigorous and independent characters are made, they will infallibly rebel against the yoke. No such person will ever feel that others have a right to control him in his concerns...and it easily comes to be considered a mark of spirit and courage to fly in the face of such usurped authority, and do with ostentation the exact opposite of what it enjoins[7]

As an example, consider the phenomenon of 'rolling coal', which involves modifying the engine of a vehicle so that it produces vastly greater quantities of particulates which appear as black clouds of exhaust. The practice was popularized in the United States as a form of anti-environmental protest in response to increasingly strict regulations on carbon emissions.[8] Note that these individuals don't simply refuse to comply with the relevant regulations. Instead, they spend significant resources in order to make a *larger* contribution to the problem than they did prior to the introduction of the regulations - that is, they "do with ostentation the exact opposite of what [the state] enjoins".

Individuals may push back simply because they don't like to be told what to do or value. They may rebel against being imposed upon by others, and not necessarily against what they are being enjoined to do or care about. Perhaps with a more nuanced application of rational methods, one that signals a willingness to recognize their concerns and engage with them in good faith, they could be brought to endorse the relevant aims. In such cases, manipulative methods (and therefore transformative manipulation) cannot be justified. But this is not an accurate description of the motives of strongly unreasonable persons. Efforts to affect a change to their normative commitments that render them conducive to practical reasonability are liable to be met with hostility in virtue of the *substance* of the change. No amount of finessing of rational methods can obviate this fact. If this is correct, then manipulative methods are still on the table.

A second reason we should doubt the efficacy of rational methods in this context is that strongly unreasonable persons are liable to regard the state's efforts as hypocritical. From the liberal perspective, the state is to regard persons as having unconditional value *qua* persons, and therefore as warranting

[7] Mill 2003, p. 146.
[8] Weigel 2014

recognition respect. This grounds its duty to maintain the stability of political conditions under which persons are reasonably able (but not required) to exercise basic autonomy. However, from the perspective of strongly unreasonable persons, being subjected to transformative measures appears to violate the state's own commitment to recognition respect. By attempting to affect a change to the target's normative commitments, the state may be perceived by the target to be interfering with their ability to exercise basic autonomy as they see fit. This is the opposite of what recognition respect seems to demand of the former. To be sure, strongly unreasonable persons do not think that persons *actually* warrant recognition respect (since they reject the idea that persons are unconditionally valuable *qua* persons). But the fact that the state appears to be violating its own commitments, and therefore evincing practical unreasonability, could easily undercut its case for adopting commitments that are consistent with practical reasonability in the eyes of those who already reject such commitments. The (successful) use of manipulative rather than rational methods would mitigate this issue, for if strongly unreasonable persons are not aware that they are being subjected to transformative measures, then the charge of hypocrisy cannot arise.

A final worry about rational methods here is that their use may undercut any effort to induce practical reasonability by way of inducing doxastic reasonability - in other words, to affect a transformation from strong unreasonability to strong reasonability. I return to the significance of this issue in the next section, but for now I will simply try to demonstrate that it is genuine worry. To see why, we need to take a short detour to consider the relationship between self-respect and respect for others.[9]

In Chapter 3 we noted that respect for others involves adopting attitudes and beliefs that are appropriate given their value, and conducting oneself in a manner that is consistent with the possibility of this value being realized.[10] Recognition respect follows from the belief that persons have unconditional value *qua* persons, while appraisal respect follows from a positive assessment of their character or conduct. Self-respect can be understood in these same terms, only directed towards one's self, i.e. having the right kinds of attitudes and beliefs about one's own value, and conducting oneself in a manner that is consistent with the possibility of this value being realized.[11] To possess

[9] There is a voluminous literature on the concept of self-respect. I make no attempt to survey it here, as this would take us far beyond the present discussion. For an excellent collection of essays on different dimensions of the topic, including its relation to respect for others, see Dillon 1995a.

[10] See §3.2.4.

[11] Dillon 1995b, 45 n31 argues that we should "...exercise caution in relying on the concept

recognition self-respect means deriving a sense of self-worth from one's status as a person as such, while appraisal self-respect means deriving a sense of self-worth from one's conduct or character.[12]

Self-respect and respect for others are connected in two ways. First, the degree to which one evinces self-respect is sensitive to the degree to which one feels respected by others.[13] Approbation from others, particularly those whose opinions matter to us, appears to us as a response to perceived value. One might doubt the veracity of their judgement, as in the case of imposter syndrome; however, when the response is more or less consistent, we cannot help but internalize it to at least some degree. Thus, to be treated with recognition respect by others is conducive to the development of recognition self-respect. Similarly, being subject to appraisal respect is conducive to the development of appraisal self-respect - indeed, exposure to *excessive* appraisal respect can lead to pathologically inflated appraisal self-respect. It is not uncommon for those who are consistently flattered for their wealth to develop a sense of self-worth that is defined almost entirely in terms of their status as a wealthy individual. Of course, this dynamic cuts both ways. Being subject to disapproval, ridicule, or even apathy from others, particularly those whose opinions we care about, can easily appear to us as a response to a perceived lack of value. If we are sufficiently secure in ourselves, then perhaps this treatment will have no impact on our sense of self-worth. But as in the case of approbation, consistent exposure to what we regard as a denial of one's value can lead one to internalize this denial, which is corrosive to self-respect in the relevant sense.[14]

The second connection is the sensitivity of respect for others to self-respect.[15] This is a psychological rather than conceptual or moral claim. Clearly, it would be too strong to say that it is impossible to genuinely respect others if one does not already have a commensurate degree of self-respect. Certainly, we very often respect others on the basis of their character or conduct precisely because they are or have achieved something that we aspire to be or do. For example, I can hold a novelist in high regard in virtue of their

of respect to make sense of self-respect". I don't quite agree with her reasoning, but leave this quarrel aside.

[12] For discussions of recognition and appraisal self-respect, see Bird 2008, Dillon 1995b, and Middleton 2006.

[13] See Cureton 2013; Dillon 1992; Hill Jr. 1991, Ch.1; Middleton 2006; Rawls 1999, pp. 155–156.

[14] For example, constant criticism of one's abilities as a musician would for many people eventually erode their belief in their having any value as a musician.

[15] This relationship is of course a central component in Kant's account of duties to self and others. See Cureton 2013 for a useful discussion on this topic.

skill despite my complete lack of literary talent, or indeed precisely *because* I lack their talent. But the connection is surely much closer in the case of recognition respect. It is difficult to believe that it is possible to genuinely value another's humanity (or at least to the appropriate degree) while simultaneously denying the value of one's own.[16] I may be able to conduct myself in a manner that is consistent with the possible realization of another's value *qua* person, but it's not clear how I could rationally hold attitudes and beliefs that are appropriate to their value without also including myself within the scope of these beliefs and attitudes (unless I am prepared to deny my own humanity).

The use of rational methods to affect a transformation from strong unreasonability to strong reasonability is liable to be self-defeating due to the mutual sensitivity of respect and self-respect. Recall that the application of rational methods renders the existence and purpose of the policy means transparent to the target. This means that strongly unreasonable persons will understand that the state intends to affect a change to their normative commitments such that they are consistent with respect for persons as such. But this treatment appears to signal beliefs and attitudes that are inconsistent with the idea of persons as having unconditional value whose realization is tied to their ability to exercise basic autonomy. The state's efforts therefore risk being perceived by the targets as evincing a failure of recognition respect. To be sure, strongly unreasonable persons already reject respect for persons *qua* persons. However, the point is that if self-respect is sensitive to respect from others, and strongly unreasonable persons interpret the state's efforts as failing to evince recognition respect for them, then these efforts will not be conducive to the development of recognition self-respect. Furthermore, since respect for others is sensitive to self-respect, and the application of rational methods in this context is not conducive to the development of recognition self-respect in strongly unreasonable persons, then it is also not conducive to their coming to evince practical reasonability that is motivated by a commitment to recognition respect, i.e. becoming strongly reasonable.

In summary, we have several reasons to doubt the efficacy of rational methods to affect a transformation to the normative commitments of strongly unreasonable persons. Their awareness of the state's efforts and the rationale

[16] Hill Jr. 1991, pp. 19–24 argues that it is not inconceivable that one might lack self-respect in the sense of regarding one's interests and plans as worthless and yet still be capable of respecting others. I'm inclined to interpret this to suggest that we can lack appraisal self-respect while still being capable of recognition respect for others, and not that the latter is possible without recognition self-respect.

are liable to provoke a backlash based on resentment at being told what to do or value, or be disregarded as an exercise in hypocrisy on the part of the state, or undercut the development of recognition self-respect. Thus, given the necessity of transformative measures to address the threat that such persons pose to stability, there is a strong case for the justifiability of manipulative methods. In other words, for the use of transformative manipulation.

5.1.2 Weakly unreasonable persons

Recall that weakly unreasonable persons are willing to comply with liberal political conditions despite endorsing a comprehensive doctrine that rejects respect for persons as such. As an example, consider the Amish, whom Jeff Spinner describes as

> ...[frowning] upon the human abilities to be self-critical and reflective; they are not interested in debate and discussion; they are not eager to experiment or sculpt their own identities...[they] escape the political demands of liberal citizenship[17]

That they reject the idea of respect for persons as such is clear from their denial of the conditions under which this value is (in part) realized, i.e. being reasonably able to exercise basic autonomy. And while they do not actively support the liberal political order, their 'escape from political demands of liberal citizenship' consists in a form of life that can nevertheless be tolerated under such an order. In this sense, they comply with its rules, and so are practically reasonable.

When do we have reason to believe that such persons pose a genuine threat to stability? Unlike in the case of strongly unreasonable persons, numbers matter. As noted in the previous chapter, what makes weakly unreasonable persons a potential threat is the contingency of their practical reasonability. Individuals and small groups of weakly unreasonable persons are willing to refrain from attempting to reshape the political order in the image of their comprehensive doctrine *when and because* they make up a small part of the wider population. They are practically reasonable in virtue of their minority status. Insofar as their numbers remain below a certain threshold, there is no reason to believe that such persons or groups pose a threat to stability. But this is not so once the threshold is passed. Certainly, it's not *inconceivable* that they might continue to support a liberal political

[17] Spinner 1996, p. 94.

order despite after gaining a sufficiently large demographic footprint. However, as discussed in §4.3.3, there are good reasons to believe that this is highly unlikely given the nature of doxastic unreasonability.

What can we say about the threshold itself? I'm sceptical that any principled specification is possible.[18] The line at which a group of weakly unreasonable persons becomes sufficiently numerous to pose a threat to stability will almost certainly depend on a variety of context-sensitive factors, e.g. the group's perception of its own power, the willingness of other groups to resist attempts to undermine the political order, the technologies that groups have access to, etc. It is simply not possible to provide a neat criterion that takes all of this complexity into account. This is all to say that determining whether a group of weakly unreasonable persons is sufficiently numerous to pose a threat to stability will sometimes require an exercise of judgement on the part of policymakers.

Why transformation?

Let us assume that a group of weakly unreasonable persons has hit the threshold at which they pose a threat to stability. Why should we doubt the efficacy of conservative measures to address this threat? One option is the use of conservative coercion - that is, the application of instruments that raise the costs of certain actions as defined by the target's own conception of self-interest. Sedition laws might serve this purpose. Certainly, such an approach could be effective, but only if the group's efforts to undermine liberal political conditions take extrajudicial forms, e.g. acts of political violence or intimidation, refusal to carry out one's duties as a public office holder, etc. History offers clear examples of weakly unreasonable groups who have undermined or overthrown a liberal political order through entirely legal mechanisms, however. Undoubtedly the most infamous example is the German elections of 1933 which produced a majority coalition headed by the Nazi party, and the subsequent passage of the Enabling Act that effectively ended the Weimar Republic.[19] Insofar as a group of weakly unreasonable persons are prepared to pursue their illiberal aims through legal means, then conservative coercion will have no purchase on them.

The use of conservative persuasion - i.e. explicit appeal to the target's existing normative commitments as giving them sufficient reason to ϕ - faces a different problem. On the one hand, weakly unreasonable persons are

[18] In this I agree with Quong 2011, p. 304.

[19] Of course, with the passage of the Enabling Act, the Nazi party shifted to strong unreasonability.

doxastically unreasonable, so appealing to moral considerations that favour practical reasonability is unlikely to move them. In other words, one would be hard-pressed to convince them to maintain their support for liberal political conditions *because* these conditions embody respect for persons as such. Their motivation for practical reasonability was prudential from the start. For the same reason, appealing to prudential considerations is unlikely to be effective, at least in every case. All else being equal, the greater the numbers and influence of a group of weakly unreasonable persons relative to that of the wider population, the less convincing will be the case that it is still in their self-interest to refrain from attempting to reshape the political order in the image of their comprehensive doctrine.

Thus, we have reason to doubt the efficacy of conservative measures to maintain the practical reasonability of groups of weakly unreasonable persons once their numbers reach a certain threshold. I do not claim that we can expect such measures to be ineffective in all cases, only that we can expect there to be cases where they are ineffective. And it is with respect to these cases that transformative measures are justified.

Manipulation and restraint

When a group of weakly unreasonable persons poses a threat to stability, it is because they are plausibly able to undermine a liberal political order in virtue of their relative numbers. However, they need not actually be engaged in such an effort to qualify as such. It is enough that they both reject the moral foundations of that order and possess the ability to plausibly impact its proper functioning. The contingent exercise of restraint on their part makes them no less a threat to stability. This fact is an important consideration in the rationale for utilizing manipulative rather than rational methods when attempting to affect a transformation in weakly unreasonable persons that preserves practical reasonability.

As noted previously, weak unreasonability is characteristic of parties for whom liberal political conditions constitute a *modus vivendi*.[20] They are willing to comply with its rules because (a) it means protection from those who would otherwise impose their beliefs on others, and (b) they cannot or do not wish to impose their beliefs on others. Absent the satisfaction of these conditions, they have no motivation for practical reasonability. Now suppose that the state utilizes rational methods in an attempt to affect a change to the normative commitments of a group of weakly unreasonable persons who

[20] See §4.3.3

pose a threat to stability. Recall that the defining feature of such methods is that the existence and rationale for the application of some instrument is transparent to the target. This means that the group will recognize and understand the significance of the state's efforts - namely, as an attempt to change their normative commitments. But this is an imposition of the kind that their practical reasonability was meant to protect them from in the first place. Therefore, condition (a) is no longer satisfied. And since it is no longer implausible for them to attempt to impose their doctrine on others, and they have sufficient motivation to do so in light of the state's actions, condition (b) also fails to obtain. Ironically, then, the state's use of rational methods to preserve the practical reasonability of weakly unreasonable persons who pose a genuine threat to stability runs the risk of *actualizing* this threat. These considerations strongly suggest that in cases where the use of transformative measures is justified, so too are manipulative methods - that is to say, we have good reason to believe that the use of transformative manipulation can be justified in certain cases of weak unreasonability.

There is a potential worry with this line of reasoning, however. If rational methods should be avoided in this context because awareness of the state's transformative efforts is liable to produce the opposite of the intended effect, then surely we face the same risks when utilizing manipulative methods. Keeping the existence or aims of policy means concealed from the target population will always be a challenge. If there is a risk that targets will abandon practical reasonability when the state is *transparent* about its efforts, then surely the discovery of surreptitious efforts will provide the targets with even stronger motivation to abandon it. Why, then, does concern about this risk not have equal force in the evaluation of manipulative methods?

The reply to this objection is fairly straightforward. The use of manipulative methods is at least capable of reducing the risk of the target becoming cognisant of the state's transformative efforts. The same cannot be said of the alternative, for the target's awareness is a constitutive feature of rational methods. That the use of either kind of method risks catalysing the very thing that the transformative measures are designed to prevent - i.e. the erosion of practical reasonability - does not mean we should draw the same conclusions from this fact. In the case of rational methods, it necessarily counts as a strike against their use. In the case of manipulative methods, it is an impetus for better policy design. Insofar as effective policy design is within reach, then manipulative methods should be favoured in certain cases of weak unreasonability.

5.1.3 Weakly reasonable persons

Recall that weakly reasonable persons endorse a comprehensive doctrine that is consistent with respect for persons as such, but favour political principles that do not reflect this principle. In other words, they are doxastically reasonable but practically unreasonable.[21] We can make sense of their opposition to a liberal order as motivated by the (mistaken) belief that their preferred illiberal order would do a better job of securing conditions under which persons are reasonably able to exercise basic autonomy. For example, one might endorse a flavour of luck-egalitarianism that prescribes state intervention to rectify any and all inequalities that are not wholly the result of individual choice (e.g. congenital talents), and this in the belief that any such inequalities impair a person's being reasonably able to exercise basic autonomy.

When do such persons pose a genuine threat to stability? As in the case of strongly unreasonable persons, individuals or small groups of weakly reasonable persons may possess the necessary resources and connections to undermine a liberal political order. However, they are unlikely to seek to achieve their aims through extrajudicial measures (e.g. political violence), for this is not consistent with their commitment to respect for persons as such. It is more plausible that they will pursue their aims through legal (or at least quasi-legal) channels, e.g. via political donations that buy legislative influence, or control over mass media organizations, or funding of partisan think-tanks, etc. On the other hand, those who lack these resources pose a genuine threat to stability when their numbers allow them to exert electoral influence. Thus, as in the case of strongly unreasonable persons, there is no necessary connection between the prevalence of weakly reasonable persons and whether they pose a genuine threat to stability.

Why transformation?

The claim that the use of transformative measures against weakly reasonable persons can be justified is somewhat counter-intuitive. The fact that they are doxastically reasonable, and so endorse the principle of respect, suggests that the state need only draw the target's attention to the inconsistency between their preferred political principles and their commitment to respect for persons. It may be challenging to get them to see this, but assuming they are genuinely committed to the principle of respect, then there is reason to be optimistic about the chances of success. If this optimism is not misplaced, then conservative measures should be sufficient for addressing a threat that

[21] See §4.3.2

weakly reasonable persons pose to stability. Therefore, transformative measures are off the table.

The problem with this argument is the tacit assumption that the inconsistency between the weakly reasonable person's comprehensive doctrine and their favoured political principles must reflect a failure to abide by their own standards of doxastic justification, i.e. what counts as good reasons for belief. But this isn't necessarily so. Suppose that A is doxastically reasonable, but favours replacing the liberal political order with a strongly communitarian one in the belief that it would be more conducive to basic autonomy. For the sake of argument, assume that she is mistaken. Perhaps such an order would inhibit individual initiative and self-development, and so would be inconsistent with respect for persons. On the one hand, she *could* have arrived at her conclusion by failing to live up to her own standards of doxastic justification, e.g. A thinks that general claims about human psychology should be grounded in sound empirical research, but for reasons of wishful thinking fails to adequately apply this standard to her own claims about workable political principles. But it is also possible that A arrives at her conclusion precisely because it satisfies what she counts as good reasons for belief. Perhaps she thinks that her willingness to subsume her own interests for the collective good is reason to expect that others will be similarly willing; or she gives more weight the claims of those who share her ideological commitments, etc.

Why does this difference matter? The success of conservative measures depends on appealing to the right normative commitments (e.g. respect for persons) *and* in a way that the target regards as providing good reasons for coming to the desired conclusion. If weakly reasonable persons pose a threat to stability in virtue of their failure to live up to their own standards of doxastic justification, then conservative measures can get purchase. It simply involves showing what they are already committed to in both moral and epistemic terms. But if they pose a threat in virtue of *satisfying* standards of what they take to be good reasons for belief, then the use of transformative measures faces a dilemma: if these measures make use of the target's standards of justification then they do not support the desired conclusion; but if they employ standards of justification that do support the desired conclusion, then they will fail to convince the target to come to it. We therefore have plausible reason to doubt the efficacy of conservative measures in at least some cases of weakly reasonable persons who pose a threat to stability.

The upshot is this: the use of transformative measures can be justified against weakly reasonable persons who pose a threat to stability if (a) they support political principles that are inconsistent with respect for persons as such, and (b) their endorsement of these principles is supported by what they take to be good reasons. The goal is to induce practical reasonability

by affecting a change to that part of the target's normative commitments that determine what counts as good reasons for belief. To return to the example above, this could mean bringing A to reject anecdotal evidence as providing sufficient warrant for general claims, or the incorporation of ideological considerations in the weighting of testimony. However, if weakly reasonable persons pose a threat to stability in virtue of (a) but not (b), then the use of transformative measures is not consistent with respect for persons.

Manipulation and epistemic backlash

The case for utilizing manipulative methods to affect the requisite transformation to the target's epistemic commitments is similar to that discussed in relation to strongly unreasonable persons.[22] When using rational methods it is to be expected that the relevant target population will be aware of the application of the policy instrument (coercive or persuasive) *as* a policy instrument, and understand the aims of the policy of which it is the means. We previously discussed the worry that strongly unreasonable persons may push back against rational transformative measures because they resent being told by others what they should, but do not currently, value. There is an analogous worry when it comes to weakly reasonable persons, namely, that they will push back against rational transformative measures that target their epistemic commitments simply because they resent being told by others how they should, but do not currently, reason. And as before, a nuanced and good faith application of rational methods may mitigate this backlash in many cases. But there are bound to be cases where it does not, particularly when a weakly reasonable person's standards of doxastic justification are closely bound up with cultural or religious practices, e.g. the centrality of faith, appeal to tradition, deference to authority. No matter how nuanced, the use of rational transformative measures to affect a change to the target's epistemic commitments in these cases may be experienced as a request to indirectly alienate themselves from important sources of meaning in their life. Where we have reason to expect this to be the case, manipulative methods can be justified.

Even if one rejects this argument, there is additional reason to think that manipulative methods can be justified in the present context. In §5.1.2 we discussed the worry that utilizing transformative measures against weakly unreasonable persons who pose a threat to stability may serve to actualize rather than mitigate the threat. The same concern crops up here, though for

[22] §5.1.1

different reasons. Note that doxastic reasonability includes the belief that persons should not only be afforded the greatest possible leeway to come to their own conclusions about what is of value in life, but also what counts as good reasons for belief. Epistemic commitments must therefore figure into the definition of recognition respect. Weakly reasonable persons are by definition doxastically reasonable, and so endorse a comprehensive doctrine that reflects these claims. It follows that any effort by the state to utilize rational methods to affect a change to the epistemic commitments of weakly reasonable persons who pose a threat to stability risks being interpreted by the targets as a failure of recognition respect. Since they already reject liberal political principles as the most effective way of securing conditions under which persons are reasonably able to exercise basic autonomy, the perceived failure of recognition respect on the part of the state risks deepening rather than allaying their scepticism. If this is plausible, then as in the case of weakly unreasonable persons, the use of rational methods against weakly reasonable persons who pose a threat to stability risks actualizing the threat rather than mitigating it. In cases where concern about this risk is warranted, the use of transformative methods - and therefore transformative manipulation - is justifiable on grounds of respect for persons.[23]

5.2 The Convergence on Strong Reasonability

The previous section identified three cases in which the use of transformative manipulation can be justified to address genuine threats to stability. An important question remains, however. What sort of transformation should policymakers aim to affect in each of the three cases? The intuitive answer is strong reasonability. I think this is correct. But this conjecture appears to be in tension with the fundamental constraint on the use transformative manipulation that was highlighted in the introduction to this chapter, namely, that the magnitude of the affected change does not exceed what is strictly necessary to mitigate the relevant threat to stability. Call this the *Minimization Principle* (MP). In this final section, I would like to briefly illustrate the apparent tension, and why it is ultimately illusory. The upshot is that the use of transformative manipulation against persons who pose a

[23] One might raise the same objection to this argument as was raised against the justification of manipulative methods in the case of weakly unreasonable persons. However, my rebuttal there applies here as well, so I will not repeat it.

genuine threat to stability is justified only if it is meant to affect a change to their normative commitments that renders them strongly reasonable.

MP refers to magnitudes of change to a target's normative commitments. It is possible to analyse these magnitudes at a very fine-grained level, e.g. the normative distance between new commitments and those they replace, the ratio of the target's commitments that have changed, the weight of the new commitments, etc. For the sake of tractability, however, I analyse magnitudes of change in terms of the relative distance between the four categories of persons. We can represent these distances via the following orderings:

1. WR | SR < WU < SU
2. WU | SU < WR < SR
3. SU | WU < WR < SR[24]

Ordering (1) states that the transition from weak reasonability (WR) to strong reasonability (SR) is of lesser magnitude than to weak unreasonability (WU) which is lesser magnitude than to strong unreasonability (SU). Orderings (2) and (3) make the analogous claims for weak unreasonability and strong unreasonability respectively.[25] These orderings reflect an assumption that should be highlighted for the sake of transparency, though I don't believe it requires much argumentation. In simple terms, the assumption is that, all else being equal, a change from doxastic reasonability to unreasonability (or vice versa) is always of greater magnitude than from practical reasonability to unreasonability (or vice versa). Or, more succinctly:

$$(DR||DU) > (PR||PU)$$

From this it follows that, for example, a transformation from weak unreasonability to strong unreasonability is of lesser magnitude than to weak reasonability (ordering (2)). This is because the former describes an expansion of an individual's application of their normative commitments (from private sphere only to public and private), while the latter describes a fundamental change to their normative commitments, i.e. from rejecting to endorsing respect for persons *qua* persons. My assumption is that the latter marks a much greater change to who they are as a person than the former.

We are now in a position to clarify the tension between the initial intuition about the aims of transformative manipulation and MP. The intuition was

[24] This is of course implied by (2), but I include it here for the sake of completeness.

[25] I omit strong reasonability since such persons are not candidates for the use of transformative manipulation.

that any justified use of such means must aim at strong reasonability, whereas MP requires affecting only so great a change as is required to address the threat to stability. The intuition and the principle are clearly consistent in the case of ordering (1). The magnitude of the transformation from weak to strong reasonability is not only smaller than to either weak or strong unreasonability, but the latter two options would do nothing to meaningfully address any threat that weakly reasonable persons pose to stability. Thus, there is no tension here. It is with respect to orderings (2) and (3) that the tension emerges. In both cases, a transformation to strong reasonability is of greater magnitude than the alternatives. Unless there are compelling reasons to believe that the alternatives would not meaningfully address the relevant threat to stability, then there will be cases where MP prescribes an aim other than strong reasonability. In fact, we do have compelling reasons to reject the alternatives.

I begin with ordering (2). On the one hand, if weakly unreasonable persons pose a genuine threat to stability then clearly the problem is not addressed by affecting a transformation to *strong* unreasonability. So MP does not require this. But why does the principle not prescribe a transformation to weak reasonability? Put simply, because doing so would very likely exacerbate the destabilizing influence of the target population. Affecting a transformation from weak unreasonability into weak reasonability would decrease the population of those who evince at least *contingent* practical reasonability, while simultaneously increasing the population of persons who evince robust practical unreasonability. The result would be a greater proportion of the population who favour a fundamentally different kind of political order and are prepared to work within the system to bring it about. Thus, although the magnitude of a transformation from weak unreasonability to weak reasonability is smaller than that to strong reasonability, the former would likely fail to meaningfully address the threat. Therefore, MP prescribes aiming at strong reasonability.

Turning to ordering (3), the magnitude of the leap from strong unreasonability to strong reasonability is larger than either alternative. Further, it would appear that any reduction in the numbers of strongly unreasonable persons would produce a corresponding increase in stability. Taken together, these facts appear to suggest that affecting a transformation from strong unreasonability to strong reasonability violates MP, since the alternatives are closer and would reduce the threat. This is not so. Recall that whether weakly unreasonable persons pose a genuine threat to stability depends on

their prevalence.[26] Affecting a transformation from strong to weak unreasonability may address the threat posed by the former *qua* strongly unreasonable persons, but at the cost of either pushing the population of weakly unreasonable persons closer to the threshold at which they become a genuine threat to stability, or increasing the magnitude of the threat if the threshold has already been crossed. Essentially, it means increasing the number of people who comply with liberal political conditions as a *modus vivendi*, and thereby weakening their motivation for continuing to do so.

Similar worries arise for transformations from strong unreasonability to weak reasonability. In this case, the target of the policy means internalises normative commitments that include the principle of respect for persons, but retain a commitment to overturning the liberal political order. Only the rationale changes. Ex-ante, it is motivated by commitments that reject respect for persons. Ex-post, it is motivated by the conviction that liberal political conditions do not do enough to ensure recognition respect for everyone. Swelling the ranks of the weakly reasonable is unlikely to reduce the threat to stability, only transform its character. Thus, although the leap from strong unreasonability to strong reasonability is of a greater magnitude than the alternatives, it is consistent with MP because these alternatives fail to meaningfully address the threat posed by strongly unreasonable persons.

In summary, the tension between MP and the intuition that transformative manipulation should always aim at strong reasonability is a merely apparent one. We have reason to expect that any use of such means to affect changes that result in strong or weak unreasonability, or weak reasonability will do nothing to address the relevant threat, or simply exacerbate it. Since this amounts to a failure on the part of the state to uphold its duty to maintain conditions under which persons are reasonably able to exercise basic autonomy, aiming at anything other than strong reasonability constitutes a failure of recognition respect, and so is impermissible.

5.3 Conclusion

My primary aim in this chapter has been to establish that there are realistic cases wherein the use of transformative manipulation to address threats to stability can be justified. In particular, I have argued that strongly unreasonable, weakly unreasonable, and weakly reasonable persons can pose a genuine threat to stability, though each for different reasons. I have at-

[26] §5.1.2

tempted to show that there are good reasons to expect that there will be cases in each category that cannot be effectively addressed through the use of either conservative measures nor rational methods. If this is correct, then in order to discharge its duty to maintain the stability of conditions under which persons are reasonably able to exercise basic autonomy, the state must make use of transformative manipulation. Where this is true, the use of transformative manipulation as policy means is an expression of recognition respect, and this despite failing to engage with persons *as* persons. It follows that even the *de facto* interpretation of APT is false.

Chapter 6

Concluding Remarks

This thesis has concerned the question of whether the liberal tradition can accommodate the use of transformative manipulation as policy means. According to the strongest version of the Absolute Prohibition Thesis (APT), the use of transformative manipulation entails a breach of the state's unconditional duty of respect for persons, and so is impermissible in principle. I have argued that this claim implicitly assumes that policy means that fail - as transformative manipulation does - to engage with persons *as* persons in at least some way are necessarily disrespectful of them. But if discharging its duty of respect requires the state to maintain the stability of conditions under which the unconditional value of persons is realized, and this necessarily includes their being reasonably able to exercise basic autonomy, then policy means that fail to engage with persons as persons are justified where their use is necessary to achieve this aim. Indeed, under these circumstances such treatment is an expression of respect for persons, even those who are subject to it. This is enough to demonstrate that APT is incorrect at least as a conceptual claim. The *de facto* interpretation of APT concedes this point, but asserts that in fact there are no cases where the state's use of transformative manipulation is plausibly necessary to maintain stability. My argument against this weaker version of APT analyses the threat that persons may pose to stability in terms of their status as strongly reasonable, weakly reasonable, weakly unreasonable, or strongly unreasonable. I have argued that each of the latter three categories contain plausible cases wherein the use of transformative manipulation is necessary to address the threat that those persons pose to stability. If this is correct, then even the *de facto* interpretation of APT must also be abandoned. Therefore, there is no in-principle incompatibility between transformative manipulation and liberal political morality.

My account makes at least three contributions to philosophical debate

about the moral limits of state action. First, the mechanisms of state action are typically described wholly in terms of persuasion, coercion, and/or manipulation. While there is nothing wrong with asking if and when states are permitted to make use of these things, it has tended to encourage an oversimplified picture of the channels through which states can exert influence over populations. The taxonomy of features of policy means discussed in Chapter 1 - i.e. instrument, method, mode, and content - provides a more nuanced picture of state power. By disentangling these features, this taxonomy can draw our attention to morally salient dimensions of policy means that tend to be overlooked, and in doing so help us to avoid conflating distinct issues when evaluating the moral permissibility of state action.

Second, my account serves to highlight an important gap in current discussions about transformative experience. Although there has been a groundswell of work on its epistemic and moral salience, very little attention has been paid to its political significance. My discussion in Chapter 1 highlights that what little discussion does exist fails to acknowledge that there are questions about the political morality of transformative experience that cannot be reduced to questions of interpersonal morality. Since transformative policy means appear to have advantages over alternatives under certain circumstances, particularly in terms of efficacy and efficiency, it is important that we gain a better understanding of the conditions (if any) under which its use can be justified.

By giving us good reason to believe that the use of transformative manipulation can be reconciled with the liberal principle of respect for persons, my account also suggests that liberalism's basic political commitments are consistent with a broader array of policy means than has traditionally been thought. This is the third contribution. As noted in Chapter 3, some liberal theorists are are prepared to accept policy means in the transformative mode insofar as they employ rational methods, while others are prepared to accommodate the use of manipulative methods in the conservative mode. But it's not clear *any* liberals are prepared to accept policy means that utilize manipulative methods in the transformative mode. I hope to have shown that while such a restriction is consistent with any plausible conception of liberal political morality, is not a necessary feature of *every* plausible conception of liberal political morality.

These results raise a number of questions that warrant further investigation. I have argued that where the use of transformative manipulation is justified, it is in virtue of the state's respect-based duty of stability. But is this the only possible justification? Surely realizing the unconditional value of persons involves more than being reasonably able to exercise basic autonomy. Perhaps certain ways of life are so antithetical to well-being that they

can only count as preventing the unconditional value of persons from being realized. If the state is justified in discouraging such ways of life, then we might wonder if transformative manipulation has advantages over alternative approaches that warrant its use here. In general, there is value in exploring whether the legitimate scope of transformative manipulation in liberal states is broader than has been argued here, or, conversely, if reasons of stability are the firm exception beyond which the utilization of such policy means can only tip the state into illiberalism.

This thesis has focused on the legitimacy of transformative manipulation as a special case of transformative policy means more generally. It is therefore only a partial treatment of a broader concern with the political morality of transformative experience. There are a number of avenues for further research here. First, we currently lack a treatment of the moral status of rationally transformative policy means. Any such account must reckon with questions raised in the literature on the epistemic and moral dimensions of transformative choice for others. In particular, in what sense rational engagement that induces a change to the target's normative commitments can *count* as rational engagement (or rational in a sufficiently strong sense) if the target cannot fully understand what life in possession of the ex-post normative commitments will truly be like. And if they cannot, does this raise worries about consent that reveal rationally transformative policy means to be less anodyne than they might appear relative to transformative manipulation? Second, we might also ask whether the moral status of rationally transformative policy means depends on their content - that is, the kinds of reasons that they are designed to bring the target to regard themselves as having to act in the desired ways. It is conceivable that there are morally salient differences between targeting a change to someone's preferences and targeting a change to their core values. Intuitively, the question posed above seems less urgent if we are talking about the former and not the latter. Furthermore, we might wonder whether the kinds of normativity expressed by different commitments is a relevant consideration. For example, all else being equal, are rationally transformative policy means that target aesthetic values less objectionable than those that target moral or epistemic values? Third, it is not clear how the choice of instrument - i.e. coercion or persuasion - affects the moral status of rationally transformative policy means. Intuitively, transformative rational coercion seems to be far more objectionable than transformative rational persuasion, but this claim requires further investigation.

There are undoubtedly a host of other issues that any reasonably complete theory of the political morality of transformative experience must address. My aim in this thesis has been to make a small contribution towards such a

theory by demonstrating that, under very specific circumstances, the use of transformative manipulation can be reconciled with the liberal commitment to respect for persons as such.

Bibliography

Abizadeh, Arash (2018). *Hobbes and the two faces of ethics*. Cambridge University Press.

Alvarez, Maria (2010). *Kinds of reasons*. Oxford University Press.

Anscombe, G. E. M. (1957). *Intention*. Oxford: Basil Blackwell.

Barandiaran, Xabier, Ezequiel Di Paolo, and Marieke Rohde (2009). "Defining agency: individuality, normativity, asymmetry, and spatio-temporality in action". In: *Adaptive Behavior* 17.5, pp. 367–386.

Barry, Brian (1995). "John Rawls and the search for stability". In: *Ethics* 105.4, pp. 874–915.

Baumgold, Deborah, ed. (2017). *Three-Text Edition of Thomas Hobbes's Political Theory: The Elements of Law, De Cive and Leviathan*. Cambridge University Press.

Benn, S. I. and W. L. Weinstein (Apr. 1971). "Being free to act, and being a free man". In: *Mind* 80.318, pp. 194–211.

Benn, Stanley I. (1990). *A theory of freedom*. Cambridge University Press.

Berlin, Isaiah (2002a). *Liberty*. Ed. by Henry Hardy. Oxford University Press.

— (2002b). "Two concepts of liberty". In: *Liberty*. Ed. by Henry Hardy. Oxford University Press. Chap. 1, pp. 166–217.

Bird, Colin (2008). "Self-respect and the Respect of Others". In: *European Journal of Philosophy* 18.1, pp. 17–40.

Bodenhamer, David J. (2005). "Probable cause". In: *The Oxford Companion to the Supreme Court of the United States*. Ed. by Kermit L. Hall. 2nd ed.

Boettcher, James W. (2004). "What is reasonableness?" In: *Philosophy & Social Criticism* 30.5–6, pp. 597–621.

Bratman, Michael E. (2007). *Structures of agency*. Oxford University Press.

Brennan, Jason (2016). *Against Democracy*. Oxford University Press.

Buchanan, Allen (2018). "Prisoners of Misbelief: The Epistemic Conditions of Freedom". In: *The Oxford Handbook of Freedom*. Ed. by David Schmidtz and Carmen E. Pavel. Oxford University Press.

Buss, Sarah (Jan. 2005). "Valuing autonomy and respecting persons: manipulation, seduction, and the basis of moral constraints". In: *Ethics* 115.2, pp. 195–235.

Cairney, Paul (2012). *Understanding public policy: theories and issues*. Palgrave Macmillan.

Carter, Ian (1999). *A measure of freedom*. Oxford University Press.

— (2008). "How are power and unfreedom related?" In: *Republicanism and Political Theory*. Ed. by Cecile Laborde and John Maynor. Blackwell Publishing, pp. 58–82.

— (2015). "Value-freeness and value-neutrality in the analysis of political concepts". In: *Oxford Studies in Political Philosophy*. Ed. by David Sobel, Peter Vallentyne, and Steven Wall. Oxford University Press.

Chang, Ruth (2013). "Incommensurability (and incomparability)". In: *The International Encyclopedia of Ethics*. Ed. by Hugh LaFollette. Blackwell Publishing Ltd, pp. 2591–2604.

Christman, John (2009). *The politics of persons*. Cambridge University Press.

Christman, John and Joel Anderson (2005). "Introduction". In: *Autonomy and the challenges to liberalism*. Ed. by John Christman and Joel Anderson. Cambridge University Press. Chap. 1, pp. 1–23.

Cohen, G. A. (2008). *Rescuing Justice and Equality*. Harvard University Press.

Connolly, William E. (1993). *The terms of political discourse*. Princeton University Press.

Coons, Christian and Michael Weber (2014). "Introduction". In: *Manipulation: Theory and Practice*. Ed. by Christian Coons and Michael Weber. Oxford University Press, pp. 1–16.

Cranor, Carl (1982). "Limitations on respect for persons theories". In: *Tulane Studies in Philosophy* 31, pp. 45–60.

Cromartie, Michael, ed. (1996). *Caesar's Coin Revisited: Christians and the Limits of Government*. Wm. B. Eerdmans Publishing Company.

Cureton, Adam (2013). "From self-respect to respect for others". In: *Pacific Philosophical Quarterly* 94, pp. 166–187.

Dagger, Richard (1997). *Civic virtues*. Oxford University Press.

Dancy, Jonathan (2000). *Practical reality*. Oxford University Press.

Darwall, Stephen (Oct. 1977). "Two kinds of respect". In: *Ethics* 88.1, pp. 36–49.

— (2006). *The Second Person Standpoint*. Harvard University Press.

Davidson, Donald (1980). *Essays on actions and events*. Oxford: Clarendon Press.

Dean, Richard (2021). "The peculiar idea of respect for a capacity". In: *Respect: Philosophical Essays*. Ed. by Richard Dean and Oliver Sensen. Oxford University Press, pp. 140–156.
Delaney, James (2006). *Rousseau and the ethics of virtue*. Continuum.
Demick, Barbara (2010). *Nothing to Envy: Real lives in North Korea*. Granata Books.
Dent, N. J. H. (2005). *Rousseau*. Routledge.
Dillon, Robin S. (1992). "How to lose your self-respect". In: *American Philosophical Quarterly* 29.2, pp. 125–139.
— ed. (1995a). *Dignity, Character, and Self-Respect*. Routledge.
— (1995b). "Introduction". In: *Dignity, Character, and Self-Respect*. Ed. by Robin S. Dillon. Routledge, pp. 1–49.
Dodds, Annaliese (2013). *Comparative public policy*. Palgrave-Macmillan.
Dolan, Paul et al. (2010). *Mindspace: influencing behaviour through public policy*. UK Cabinet Office and Institute for Government. URL: `https://www.instituteforgovernment.org.uk/sites/default/files/publications/MINDSPACE.pdf` (visited on 01/18/2024).
Douglass, Robin (2015). *Hobbes and Rousseau: nature, free-will, and the passions*. Oxford University Press.
Dworkin, Gerald (1988). *The theory and practice of autonomy*. Cambridge University Press.
Elster, Jon (1983). *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge University Press.
Faden, Ruth R., Tom L. Beauchamp, and Nancy M. P. King (1986). *A history and theory of informed consent*. Oxford University Press.
Feinberg, Joel (1986). *The moral limits of the criminal law volume III: Harm to self*. Oxford University Press.
Forst, Rainer (2013). *Tolerance in conflict: past and present*. Cambridge University Press.
Frankfurt, Harry G. (Jan. 1971). "Freedom of the will and the concept of a person". In: *The Journal of Philosophy* 68.1, pp. 5–20.
Frederick (2021). *Frederick the Great's philosophical writings*. Ed. by Avi Lifschitz. Princeton University Press.
Gaus, Gerald F. (2004). "The diversity of comprehensive liberalisms". In: *Handbook of Political Theory*. Ed. by Gerald F. Gaus and Chandran Kukathas. Sage, pp. 100–114.
— (2005). "The place of autonomy wihtin liberalism". In: *Autonomy and the Challenges to Liberalism*. Ed. by John Christman and Joel Anderson. Cambridge University Press, pp. 272–306.
Gray, Tim (1990). *Freedom*. Issues in Political Theory. MacMillian Education Ltd.

Green, T. H. (2006). "Liberal legislation and freedom of contract". In: *The Liberty Reader*. Ed. by David Miller. Routledge, pp. 21–32.
— (2011). "On the Different Senses of 'Freedom' as Applies to Will and to the Moral Progress of Man". In: *Works of Thomas Hill Green Vol 2.* Ed. by R. L. Nettleship. Cambridge University Press, pp. 307–333.
Hall, Robert King, ed. (1949). *Kokutai no hongi (Cardinal Principles of the National Entity of Japan)*. Harvard University Press.
Harris, Sam (Feb. 6, 2012). *The Pleasures of Drowning*. URL: `https://www.samharris.org/blog/the-pleasures-of-drowning` (visited on 01/15/2024).
Hill Jr., Thomas E. (1991). *Autonomy and Self-Respect*. Cambridge University Press.
Hobbes, Thomas (2003). *Leviathan*. Ed. by Richard Tuck. Revsied Student Edition. Cambridge Texts in the History of Political Thought. Cambridge University Press.
Horton, John (1994). "Three (apparent) paradoxes of toleration". In: *Synthesis Philosophica* 17, pp. 7–20.
Howard, Dana (2015). "Transforming others: on the limits of "you'll be glad you did It' reasoning". In: *Res Philosophica* 92.2, pp. 341–370.
Hurka, Thomas (1987). "Why value autonomy". In: *Social Theory and Practice* 13.3, pp. 361–382.
Jackson, Frank (1982). "Epiphenomenal qualia". In: *Philosophical Quarterly* 32, pp. 127–136.
James Martin Center for Nonproliferation Studies (Sept. 23, 2002). *"Suitcase Nukes": A Reassessment*. URL: `https://nonproliferation.org/suitcase-nukes-a-reassessment/` (visited on 02/23/2024).
Kelly, Erin and Lionel McPherson (2001). "On tolerating the unreasonable". In: *The Journal of Political Philosophy* 9.1, pp. 38–55.
Klosko, George (2004). "Theoretical foundations". In: *Democratic Procedures and Liberal Consensus*. Oxford University Press, pp. 19–41.
— (2015). "Rawls, Weithman, and the Stability of Liberal Democracy". In: *Res Publica* 21, pp. 235–249.
Korsgaard, Christine (1996). *Creating the kingdom of ends*. Cambridge University Press.
Korsgaard, Christine M. (2021). "Valuing our humanity". In: *Respect: Philosophical Essays*. Ed. by Richard Dean and Oliver Sensen. Oxford University Press, pp. 171–191.
Kramer, Matthew H. (2003). *The quality of freedom*. Oxford University Press.
— (2008). "Liberty and domination". In: *Republicanism and Political Theory*. Ed. by Cecile Laborde and John Maynor. Blackwell Publishing, pp. 31–57.

Kramer, Matthew H. (2018). "Conceptual Analysis and Distributive Justice". In: *The Oxford Handbook of Distributive Justice*. Ed. by Serena Olsaretti, pp. 367–386.

Kruglanski, Arie W., Jocelyn J. Bélanger, and Rohan Gunaratna (2019). *The three pillars of radicalization: needs, arratives, and networks*. Oxford University Press.

Kymlicka, Will (1996). "Two models of pluralism and tolerance". In: *Toleration: An Elusive Virtue*. Ed. by David Heyd. Princeton University Press, pp. 81–105.

Lambert, Enoch and John Schwenkler, eds. (2020a). *Becoming Someone New: Essays on Transformative Experience, Choice, and Change*. Oxford University Press.

— (2020b). "Editor's Introduction". In: *Becoming Someone New: Essays on Transformative Experience, Choice, and Change*. Ed. by Enoch Lambert and John Schwenkler. Oxford University Press, pp. 1–15.

Larmore, Charles (Dec. 1999). "The moral basis of political liberalism". In: *The Journal of Philosophy* 96.12, pp. 599–625.

List, Christian and Philip Pettit (2011). *Group agency*. Oxford University Press.

Locke, John (2013). *Two treatises of government*. Ed. by Peter Laslett. Cambridge University Press.

— (2016). "A letter concerning toleration". In: *Second Treatise of Government and A Letter Concerning Toleration*. Ed. by Mark Goldie. Oxford University Press, pp. 123–168.

Lowi, Theodore J. (1972). "Four systems of policy, politics, and choice". In: *Public Administration Review* 32.4, pp. 298–310.

MacCallum, Gerald C. (July 1967). "Negative and positive freedom". In: *The Philosophical Review* 76.3, pp. 312–334.

McCabe, David (2010). *Modus vivendi liberalism*. Cambridge University Press.

Middleton, David (2006). "Three Type of Self-Respect". In: *Res Publica* 12, pp. 59–76.

Mill, John Stuart (2003). *On liberty*. Ed. by David Bromwich and George Kateb. Rethinking the Western Tradition. Yale University Press.

— (2015). "Considerations On Representative Government". In: *John Stuart Mill: On Liberty, Utilitarianism and Other Essays*. Ed. by Mark Philp and Frederick Rosen. Oxford University Press, pp. 181–388.

Miller, David (1984). "Constraints on freedom". In: *Ethics* 94.1, pp. 66–86.

Mussolini, Benito and Giovanni Gentile (1932). "The Doctrine of Facism". In: Mussolini, Benito. *Facism Doctrine and Institutions*. Ardita Publishers, pp. 7–42. URL: https://ia600800.us.archive.org/14/items/

TheDoctrineOfFascismByBenitoMussolini/The%20Doctrine%20of%20Fascism%20by%20Benito%20Mussolini.pdf.
Neuhouser, Frederick (2000). *Foundations of Hegel's Social Theory*. Harvard University Press.
Neumann, Michael (2004). "Can't we all respect each other a little less?" In: *Canadian Journal of Philosophy* 34.4, pp. 463–484.
Nozick, Robert (1974). *Anarchy, state, and utopia*. Blackwell Publishers Ltd.
Numao, J. K. (2013). "Locke on Atheism". In: *History of Political Thought* 34.2, pp. 252–272.
Nussbaum, Martha (2011). "Perfectionist liberalism and political liberalism". In: *Philosophy & Public Affairs* 39.1, pp. 3–45.
O'Hagan, Timothy (1999). *Rousseau*. Routledge.
Oppenheim, Felix (1961). *Dimensions of Freedom*. St Martin's Press.
— (1985). "'Constraints on freedom' as a descriptive concept". In: *Ethics* 95.2.
Orwell, George (2021). *Nineteen eighty-four*. William Collins.
Paul, L. A. (2014). *Transformative Experience*. Oxford University Press.
Paul, L. A. and Cass R. Sunstein (Sept. 17, 2019). "'As Judged By Themselves': Transformative Experiences and Endogenous Preferences". In: URL: https://dx.doi.org/10.2139/ssrn.3455421.
Pettigrew, Richard (2023). "Nudging for changing selves". In: *Synthese* 201 (22). URL: https://doi.org/10.1007/s11229-022-04020-2.
Pettit, Philip (2002). *Republicanism: a theory of freedom and government*. Calendon Press Oxford.
— (2008). "Republican freedom: three axioms, four theorems". In: *Republicanism and Political Theory*. Ed. by Cecile Laborde and John Maynor. Blackwell Publishing, pp. 102–130.
Popper, Karl (2013). *The open societies and its enemies*. Princeton University Press.
Quong, Jonathon (2011). *Liberalism without perfection*. Oxford University Press.
Rawls, John (1999). *A theory of justice*. Belknap Press Harvard University Press.
— (2001). *Justice as fairness: a restatement*. Ed. by Erin Kelly. The Belknap Press of Harvard University Press.
— (2005). *Political liberalism*. Columbia University Press.
Raz, Joseph (1986). *The morality of freedom*. Clarendon Press - Oxford.
— (2004). *Value, respect, and attachment*. Cambridge University Press.
Romano, Claude (2004). "Eleutheria". In: *Dictionary of Untranslatables: A Philosophical Lexicon*. Ed. by Barbara Cassin. Princeton University Press, pp. 250–257.

Rousseau, Jean-Jacques (1979). *Emile*. Basic Books.
— (1993). "A discourse on political economy". In: *The Social Contract and Discourses*. Ed. by G.D.H. Cole. Everyman Library.
— (2002a). "Discourse on the Origin and the Foundations of Inequality of Mankind". In: *The Social Contract and the First and Second Discourses*. Ed. by Susan Dunn. Yale University Press, pp. 69–148.
— (2002b). "The Social Contract". In: *The Social Contract and the First and Second Discourses*. Ed. by Susan Dunn. Yale University Press, pp. 149–256.
Ryan, Alan (2007). "Liberalism". In: *A Companion to Contemporary Political Philosophy*. Ed. by Robert E. Goodin, Philip Pettit, and Thomas Pogge. Blackwell Publishing. Chap. 14, pp. 360–382.
Salamon, Lester M., ed. (1989). *Beyond privatization: the tools of government action*. The Urban Institute Press.
Scanlon, T. M. (1998). *What we owe to each other*. The Belknap Press of Harvard University Press.
Schlosser, Markus (2019). *Agency*. Ed. by Edward N. Zalta. The Stanford Encyclopedia of Philosophy. URL: `https://plato.stanford.edu/archives/win2019/entries/agency/`.
Schneider, Anne and Helen Ingram (May 1990). "Behavorial assumptions of policy tools". In: *The Journal of Politics* 52.2, pp. 510–529.
Shklar, Judith N. (1989). "The liberalism of fear". In: *Liberalism and the Moral Life*. Ed. by Nancy L. Rosenblum. Harvard University Press, pp. 21–38.
Skinner, Quentin (1998). *Liberty before Liberalism*. Cambridge University Press.
— (2008). "Freedom as the absence of arbitrary power". In: *Republicanism and Political Theory*. Ed. by Cecile Laborde and John Maynor. Blackwell Publishing, pp. 83–101.
Smith, Kevin B. and Christopher W. Larimer (2009). *The public policy theory primer*. Westview Press.
Solzhenitsyn, Aleksandr (2018). *The Gulag Archipelago*. Vintage Classics.
Spinner, Jeff (1996). *The Boundaries of Citizenship: Race, Ethnicity, and Nationality in the Liberal State*. The Johns Hopkins University Press.
Steiner, Hillel (1983). "How free: computing personal liberty". In: *Of Liberty*. Ed. by A. Phillips Griffiths. Cambridge University Press, pp. 73–90.
— (1994). *An essay on rights*. Blackwell Pubishers.
Stevens, Robert (1988). "Coercive Offers". In: *Australasian Journal of Philosophy* 66, pp. 83–95.
Stewart, Katherine (2020). *The power worshippers: inside the dangerous rise of religious nationalism*. Bloomsbury Publishing.

Thaler, Richard H. and Cass R. Sunstein (2008). *Nudge: improving decisions about health, wealth, and happiness*. Yale University Press.

Tuomela, Raimo (2013). *Social ontology: collective intentionality and group agents*. Oxford University Press.

Ullmann-Margalit, Edna (2006). "Big Decisions: Opting, Converting, Drifting". In: *Royal Institute of Philosophy Suppplement* 58, pp. 157–172.

Velleman, J. David (2000). *The possibility of practical reason*. Clarendon Press Oxford.

Verdung, Evert (2010). "Policy instruments: typlogies and theories". In: *Carrots, Sticks and Sermons: Policy Instruments and Their Evaluation*. Ed. by Marie-Louise Bemelmans-Videc, Ray C. Rist, and Evert Verdung. Comparative Policy Analysis Series. Transaction Publishers. Chap. 1, pp. 21–50.

Waldron, Jeremy (2004). "Liberalism, political and comprehensive". In: *Handbook of Political Theory*. Ed. by Gerald F. Gaus and Chandran Kukathas. Sage Publications. Chap. 7, pp. 89–99.

— (2010). *God, Locke, and Equality*. Cambridge University Press.

Wall, Stephen (2003). "Freedom as a political ideal". In: *Autonomy*. Ed. by Ellen Frankel Paul, Fred D. Miller Jr., and Jeffrey Paul. Cambridge University Press, pp. 307–334.

— (2015). "Introduction". In: *The Camridge Companion to Liberalism*. Ed. by Stephen Wall. Cambridge University Press, pp. 1–18.

Ware, Owen (2023). *Kant on Freedom*. Cambridge University Press.

Weigel, David (July 3, 2014). *Rolling Coal. Conservatives who show their annoyance with liberals, Obama, and the EPA by blowing black smoke from their trucks.* URL: `https://slate.com/news-and-politics/2014/07/rolling-coal-conservatives-who-show-their-annoyance-with-liberals-obama-and-the-epa-by-blowing-black-smoke-from-their-trucks.html` (visited on 01/18/2024).

Weithman, Paul (2010). *Why Political Liberalism? On John Rawls's Political Turn*. Oxford University Press.

Wood, Allen W. (2008). *Kantian ethics*. Cambridge University Press.

Zimmerman, David (1981). "Coercive Wage Offers". In: *Philosophy and Public Affairs* 10, pp. 121–145.