Rawls's Conception of Autonomy[*]

Anthony Taylor, University of Oxford

The aim of this chapter is to elucidate Rawls's conception of autonomy and the role it plays in his thought across *A Theory of Justice* (1971; Revised Edition 1999) and *Political Liberalism* (1993; Expanded Edition 2005). A distinctive feature of this conception is that it takes seriously the threat to individual self-governance that can arise from the ways in which we are shaped by our social and political institutions. The idea that social institutions play a significant role in shaping the motivations and self-understanding of citizens has its origins in the work of Rousseau, but commentators on Rawls's work have, I will suggest, been insufficiently attentive to the role it can play in supporting his goal of uncovering principles that would enjoy uncoerced stability in a well-ordered society.[1]

This chapter will spell out Rawls's conception of autonomy and trace its connections to wider discussions of autonomy. Since the literatures on Rawls and autonomy are both large and have developed independently, there is value in drawing these connections for future work in both areas. In addition to this, the chapter will argue for two conclusions. First, that—despite certain appearances to the contrary—Rawls has an important autonomy-based commitment that is consistent across his two main works. Second, that this commitment is not, as some have argued, unable to play a justificatory role in political liberalism. On the contrary, I suggest that Rawls's conception of autonomy motivates his aim of finding principles of justice that can be stable, and so illuminates his later commitment to a political liberalism.

[*] This is the Accepted Manuscript version of a chapter forthcoming in *The Routledge Handbook of Autonomy*. For their comments on an earlier version, I am grateful to Ben Colburn, Collis Tahzib, and Paul Weithman.

§1. Autonomy

Individual autonomy is an ideal of self-direction or self-governance. To achieve it, we must rule ourselves rather than being ruled by others. To put it in an another now common phrase, to be autonomous is to be the author of the story of our life—the person who exercises control over its shape and direction (Raz, 1986: 386). But while the idea of self-governance gives us the concept of autonomy, conceptions of autonomy offer varying accounts of what is required to for us to achieve self-rule. It is not the aim of this chapter to settle any of the major controversies as to how the idea of autonomy is best understood. However, to see how Rawls's view fits in we will need before us a picture of the main aspects of conceptions of autonomy. I therefore begin with a sketch of these.

Authenticity and Alienation

A first aspect of autonomy is what we might call authenticity or non-alienation. To be autonomous, our desires and motivations must be *our own* rather than external impositions on us. To take a familiar example, consider a person in the thralls of a nicotine addiction who resents their desire to smoke. This person might feel alienated from this desire; they might not identify with it at all and wish that they could be free of it. Intuitively, such an addiction seems to diminish the addict's autonomy in some sense.

Philosophers have tried to capture this thought in different ways, but one well-known way to do so is by appeal to a hierarchical analysis.[2] On this view, the way to analyse examples like that of the addict is to consider the person's second-order desires. We can describe this case by saying that the addict has a first-order desire to smoke, but a second-order desire (a desire about his first-order desire) to be rid of the desire to smoke. Here the addict's lack of autonomy has its source in the conflict between his first and second-order desires. One possible necessary condition for autonomy is therefore that the agent identifies with their motivations, where such identification is

marked by a coherence between their first and second-order desires. But we need not accept that condition here. What is important is that it is one way of capturing the idea that we can lack autonomy when our lives are not lived in accordance with our deep commitments. The addict who regrets his addiction lives his life in ways that conflict with his reflectively held views about how it ought to be lived, and this is ultimately the source of his failure to be fully autonomous.

Procedural and Substantive Independence

A related aspect of autonomy is that a person can fail to be autonomous due to influences on their desires that subvert their reflective capacities. If we only want what we want because we have been manipulated or indoctrinated, then our autonomy is clearly threatened. In these cases, there is a failure of what Gerald Dworkin has called *procedural independence* (1988: 18–19).

We all have our desires, attitudes, and beliefs influenced in various ways by the particular circumstances that we inhabit. Since these circumstances are not chosen by us, this raises the important question of how we can be autonomous despite being so heavily influenced by our environment. If our conception of autonomy is to be a feasible one, something that individuals can achieve, then it cannot hold that to be autonomous we must be entirely free from external influences. It is not possible for anyone to live such an unencumbered life. So, if we are to have an achievable conception of autonomy, the questions that we must answer are: what ways of influencing individuals are compatible with their autonomy? How are the distinctions between influences like "hypnotic suggestion, manipulation, coercive persuasion [and] subliminal influence" to be drawn? How are we to distinguish between education on the one hand, and indoctrination on the other (*Ibid*.)?

Though most would agree autonomy requires a form of procedural independence, a more challenging question is whether it also requires *substantive independence*. Are

there substantive limits on how we can conduct our lives consistent with maintaining our autonomy? If you choose to live a life where you simply obey someone else's orders, have you thereby forfeited your autonomy? For some, to be autonomous we must see ourselves as sovereign in deciding what to believe and what to do (Scanlon, 1972: 215). We are sovereign in this way when these decisions are up to us, when we are the one who has the last word on them (Enoch, 2017: 32). This does not mean that the autonomous person never relies on the judgment of others. They may do so, but what they must not do is to accept the judgment of others without any independent consideration. Others, however, argue for a conception of autonomy that places no substantive limits on what the autonomous person may do. If they choose to live a life following the orders of their priest, this need not involve a sacrifice of their autonomy (Dworkin, 1988: 21–33).

Moral Autonomy

A further aspect of the idea of autonomy is moral autonomy. This is the idea, which has its origins in Kant, that the moral law is self-legislated.

For Kant, our autonomy consists in our being subject only to our own wills and not to the wills of others. Therefore, autonomy does not require that we are not bound by any laws at all; it requires instead that the laws we are subject to are laws of our own making—including the moral law. Precisely what it means for the moral law to be self-legislated is a matter of debate among Kantians, and we will come shortly to the particular understanding of this idea that Rawls appeals to in *Theory*. But at the very least it means that moral principles are not given to us by God, nature, or some other external authority.

An important question for the idea of moral autonomy is how it can be squared with the further Kantian ideas that the moral law is objective, obligatory, and necessarily applies to all rational beings. How can the moral law be necessary and obligatory if

its normative force depends on our giving it to ourselves? What if we decide not to give it to ourselves?[3] Since Rawls's conception of autonomy is explicitly Kantian, we will return to these questions shortly to see what sense can be made of the idea of moral principles as self-legislated within his theory.

§2. Autonomy in *A Theory of Justice*

I will return to these various aspects of autonomy at the end of the chapter, to consider how Rawls's conception relates to them. First, however, I want to consider the role that autonomy plays across Rawls's two main works, beginning with *A Theory of Justice.*

The central and most discussed argument of Rawls's *Theory* is the argument from the original position. This argument considers what principles of justice would be chosen by rational parties behind a veil of ignorance: a position of equality in which each party is deprived of knowledge of their race, ethnicity, gender, age, income, wealth, natural endowments, comprehensive doctrine, and to which generation of history they belong (2001: 14–18). In these circumstances, Rawls argued that two lexically ordered principles justice would be chosen:

> First, a principle of equal basic liberties, holding that "each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all".
>
> Second, a principle to regulate social and economic inequality, holding that such inequalities must satisfy two conditions: "first, [being] attached to offices and positions open to all under conditions of fair equality of opportunity; and second, [being] to the greatest benefit of the least-advantaged members of society" (*Ibid*: 42–43).

These two principles, and the lexical priority of the first principle over the second, are the core of Rawls's conception of justice: Justice as Fairness.

Suppose we agree that Justice as Fairness would be chosen in the original position. What is the significance of this fact? Why should it lead us to accept these principles of justice? The force of the argument from the original position is typically understood as stemming from the idea that the constraints on the choice there are fair. On this reading, the justification of the principles is broadly based on considerations of coherence in reflective equilibrium. Because the ideals of fairness we already accept lead us to see the choice in the original position as fair, we must see the principles themselves as fair.

However, Rawls also offers a quite different cast on the argument from the original position, which he calls the Kantian interpretation. This interpretation connects Justice as Fairness to a Kantian conception of autonomy, holding that that the original position offers a "procedural interpretation of Kant's conception of autonomy and the categorical imperative" (1999: 226). It is worth quoting Rawls's explanation of the relationship between the original position and the Kantian conception of autonomy in full:

> Kant held, I believe, that a person is acting autonomously when the principles of his action are chosen by him as the most adequate possible expression of his nature as a free and equal rational being. The principles he acts upon are not adopted because of his social position or natural endowments, or in view of the particular kind of society in which he lives or the specific things that he happens to want. To act upon such principles is to act heteronomously. Now the veil of ignorance deprives the persons in the original position of the knowledge that would enable them to choose heteronomous principles. The parties arrive at their choice together as free and equal rational persons knowing only that those circumstances obtain which give rise to the need for principles of justice (*Ibid.*: 222).

The Kantian conception of autonomy, on Rawls's understanding, directs us to act on principles that express our nature as free and equal rational beings as much as is possible. To express our nature in this way we must act on the principles that we would choose to act on if our nature as persons were *the decisive determining element of*

our choice. To put it another way, to be autonomous we must act on the principles we would choose if we bracketed the various contingent features of our circumstances—our particular characteristics, the society we inhabit, our position within that society—and made our decision solely on the basis of our rational nature. And this is precisely the scenario that original position aims to model via the veil of ignorance. Since the parties in the original position do not know the various contingent features of their circumstances, they are forced to choose principles of justice by deciding based on their nature as free and equal rational beings; there is no other material available to them to make this choice.

The Kantian interpretation is the first role that a conception of autonomy plays in *Theory*: it offers a distinctive way of understanding the force of the argument from the original position. For some, this has been seen as offering a deeper and more compelling argument for Justice as Fairness than the argument from reflective equilibrium. Stephen Darwall, for example, writes that while the argument from reflective equilibrium aims to make our intuitions about justice consistent, the Kantian interpretation explains why justice is something we should care about, by embedding the principles in a broader theory of practical reason (1976: 164–165).

There is, however, a second and more crucial role that this Kantian conception of autonomy plays in *Theory*. Following the argument from the original position, Rawls seeks to show that Justice of Fairness would be stable in the conditions of a well-ordered society, or at least more stable than its chief rivals such as utilitarianism. A well-ordered society is one in which: (i) all citizens accept Justice as Fairness; (ii) its principles effectively regulate society's major social and political institutions; and (iii) these two facts are public knowledge (Rawls, 2001: 8–10). One way to stabilise the well-ordered society would be Hobbesian: we could introduce a sovereign with the power to threaten sufficient coercion to ensure that everyone would continue to comply with the principles. The stability Rawls wants to establish is different,

however. He wants to show that a society could be stabilised by its citizens freely exercising their practical reason. This requires showing that the those who grew up in the well-ordered society would come, as part of their upbringing, to have a sufficiently strong desire to act justly; or, in his words, a strong *sense of justice*. If the members of that society would have a sturdy sense of justice—if they would have a desire to act justly that tended to be effective even when they had other conflicting desires—then society could be stabilised without the Hobbesian recourse to the threat of coercion.

Rawls calls a society that is stable in this way *stable for the right reasons*. This is a society that is stable because its citizens see the principles of justice that are implemented their as congruent with their good (Freeman, 2002). We can describe such a society as one in which each citizen comes to freely and reflectively endorse the principles of justice that are applied within it. A society that is stable in this way is clearly a lofty ideal, but it is one that Rawls believes could be achieved if Justice as Fairness were perfectly implemented in favourable conditions.[4] The full argument for this conclusion draws on an empirical account of moral and psychological development, and here is not the place to recount the full details of it. What is significant for our purposes, though, is the central role that autonomy plays in this argument. In a key move, which was meant to secure the case for the stability of Justice as Fairness, Rawls supposed that every citizen of the well-ordered society would have an effective desire to express their nature as a free and equal rational being (1999: §86). Given the Kantian interpretation of the original position set out above, the presence of this desire among citizens of the well-ordered society was an important basis of social stability. Since citizens would have an effective desire to express their nature in this way, they would have an effective desire to act on the principles that would be chosen in a scenario where that nature was the decisive determining element of the choice—the original position. Therefore, if the principles of Justice as Fairness are the ones that would be chosen by the parties in the original position, then they would be stabilised by citizens' desire for Kantian autonomy.[5]

As I noted above, the role that Rawls's conception of autonomy plays in the stability argument is more crucial than the role it plays in the interpretation of the original position. When it comes to the argument from the original position, the Kantian interpretation is presented by Rawls as having the status of an optional extra—he does not suggest that readers must find this conception of autonomy compelling to accept the argument. After all, even if we reject the idea of autonomy that underlies the Kantian interpretation, we might still find the argument from the original position compelling on the basis that it systematises our considered judgments in reflective equilibrium. However, by the time we reach the stability argument in *Theory*, Rawls's conception of autonomy loses this status as an optional extra. If the defensibility of Justice as Fairness depends on its stability (for the right reasons), and if the argument for its stability depends on the supposition that everyone in the well-ordered society would have an effective desire for Kantian autonomy, then this conception of autonomy has an ineliminable role to play in the central argument of the book.

§3. Autonomy in *Political Liberalism*

The account just provided of the role of autonomy in *Theory* is not, so far as I am aware, subject to any major interpretive controversy. However, when it comes to the continuing role of autonomy in Rawls's view when it is recast as a political liberalism, we will not be so fortunate. Here I will begin by presenting what I take to be a typical way of understanding the place of autonomy in *Political Liberalism*, according to which it has a quite limited role to play.

Let us then begin, then, with the typical story. *Political Liberalism* is born of the fact that there is a serious problem with the stability argument of *Theory* (2005: xl–xli). As we have seen, that stability argument depended on the claim that everyone who grew up in the well-ordered society would come to have an effective desire for Kantian autonomy. But Rawls came to doubt this supposition, as he came to believe that any society well-ordered by liberal principles would contain a plurality of conflicting

world views (*Ibid.*: xxxvi). Given this pluralism, there would be some, perhaps many, who either lacked the desire for Kantian autonomy entirely or gave it little weight in their practical reasoning. How, then, could the well-ordered society be stabilised?

Rawls's new answer to this question appeals to the idea of an overlapping consensus (*Ibid.*: lecture IV). Such a consensus involves the citizens of the well-ordered society, who adhere to a variety of different comprehensive doctrines, each finding reasons from within those doctrines to support liberal principles. When this kind of consensus holds, the citizens of the well-ordered society do not support the principles because of their desire for Kantian autonomy, but rather for a variety of different reasons stemming from their wider doctrines. In his model case of an overlapping consensus, Rawls suggested that three views could affirm his principles of justice. The first was a religious doctrine with an account of free faith, which thereby provides a basis for toleration and basic liberties. The second was a general liberal doctrine of the sort second endorsed by Kant and John Stuart Mill. And the third was a pluralist view, the domain of value is irreducibly plural. Of these three views, only some variants of the liberal doctrine might be characterised by their acceptance of a Kantian conception of autonomy (*Ibid.*: 145).

It looks, then, like the role that autonomy ultimately plays in the statement of Justice as Fairness as a political liberalism is highly limited. Though some may accept the principles because of their relationship to the Kantian conception of autonomy, many will accept them for other reasons. The justification for those principles therefore makes no necessary reference to this conception of autonomy; one might reject the conception of autonomy entirely and still accept the argument for the principles of justice. This limited role for autonomy coheres with the views of many readers of *Political Liberalism*. Indeed, many have thought that autonomy could not play a more expansive role in the view. After all, if we want to make liberalism a doctrine that is suitable for societies that are characterised by reasonable pluralism, we will need to

jettison from its justificatory apparatus any ideas that will prove to be controversial among reasonable citizens. Since some reasonable citizens may reject any autonomy, there is no place for it in the justificatory apparatus of political liberalism (Quong, 2013: 270–271; Rostbøll, 2011: 341–342).

§4. Autonomy and Stability

I will now present an alternative picture of the role of autonomy across *Theory* and *Political Liberalism*, one that gives it an important justificatory role in the latter work. This alternative picture also starts with Rawls's requirement that a conception of justice be stable. Note that it is this requirement that seems to lead to the jettisoning of autonomy from the argument for liberalism, for in *Political Liberalism* this stability requirement is the expression of the general idea that liberal justice ought to be compatible with the range of doctrines that we are likely to find in a pluralistic liberal society. But there is an important question, which receives too little attention, about *why* liberal justice must satisfy this stability requirement. The requirement is, after all, idiosyncratic to Rawls and his followers. Most political philosophers do not think they need to show that their favoured principles of justice will be compatible with the range of views that citizens of a liberal society are likely to hold. On the contrary, many in the history of philosophy have argued that it is no objection to their preferred principles that they are unlikely to receive wide uptake, or indeed are entirely unsuitable for being widely accepted by citizens. Why then should we demand that a liberal theory of justice be acceptable to all reasonable citizens?

Numerous answers have been proposed to this question. Some have seen the stability requirement as simply a matter of making a liberal democratic theory of justice consistent. On this view, suggested by Burton Dreben, in elaborating the stability argument Rawls was engaged in "a certain kind of very complex conceptual analysis" considering the question "Is the notion of a constitutional liberal democracy internally consistent or coherent?" (2003: 322). Other defenders of the stability requirement have

seen it as justified by the need to make our theories realistically utopian or at the 'limits of practical possibility'.[6] The model of a society well-ordered by Rawls's principles of political justice is a distant ideal and a lofty aspiration, but in showing that the principles would be stably accepted in these conditions Rawls shows that this ideal is not unachievable or overly utopian (Quong, 2011: 158–160). And, finally, still other defenders of the stability requirement argue that it is supported by a general condition that any normative theory must satisfy. When we defend normative principles, we must think that it is desirable for there to be a social consensus on the set of beliefs that would lead people to act in accordance with them. And a stable well-ordered society is simply one in which there is this consensus on the beliefs needed to make people act in accordance with Rawls's principles in a pluralistic liberal society (Krasnoff, 1998: 269–292).

Though I do not have space to establish it here, I believe that each of these answers runs into serious difficulties. Instead, I think the most promising defence of the stability requirement is one that make essential reference to a conception of autonomy.[7] I will aim in what follows to set out the basic features of this autonomy-based argument for the stability requirement and go on to suggest that it plays an ineliminable role in Rawls's later thought.

To see how autonomy might be involved in the case for the stability condition, I begin by noting the considerable shaping influence that social and political institutions have over future citizens. Since the kind of education and upbringing we have is guided by the principles of social and political justice that are operative or at least dominant in our society, we have each had our character and self-conception to some degree shaped by such principles. This fact is highly significant, for as I noted above, we can fail to be autonomous if our life is to some extent the product of alien forces. An important threat to our autonomy is therefore the possibility that we might come, as we reach maturity, to reject the guiding principles that were operative on our political,

social, and educative institutions during our upbringing. If we reject these principles, and the institutions guided by them, we will be rejecting as alien a significant force that operated to make us into who we are. And if we reject *this* then we will, to some extent, be rejecting our character and self-conception, at least insofar as it is the product of these forces.

Rawls couches exactly this point in terms of the attitudes that citizens of a well-ordered society might have toward their sense of justice. As these citizens grow up, they will come, Rawls supposes, to have an effective desire to act justly – an effective sense of justice. But some citizens may come to have doubts about their sense of justice. Knowing that they have grown up in a society that aims, via its educative institutions and norms about the upbringing of children, to inculcate the desire to act justly in its future citizens, why should they accept the desires stemming from it as having rational authority? Why should they not see them instead as a kind of emotional technology implanted into them against their will (Rawls, 1999: 451–452)? The stability argument is a way of raising the question that the citizens of the well-ordered society might come to reject their sense of justice. If the principles of justice failed to be stable, then some citizens would see their sense of justice as something they should aim to ignore or rid themselves of. They would see their desire to act justly in this way even when they felt it having a strong pull over them. (Think, for example, of the phenomenon of 'Catholic guilt' among those who were raised Catholic but are now atheists). By contrast, if the principles of justice are stable, then each citizen will seem themselves as having reasons to accept and maintain their sense of justice going forward.

When Rawls's stability requirement is satisfied, then, each citizen of the well-order society is autonomous in the sense that they experience their sense of justice as a product of their own will rather than as an alien imposition. The conception of autonomy at work here is what Rawls calls the *political value of full autonomy*. This is a

value that he claims is realized by the citizens of a well-ordered society "in their recognition and informed application of the principles of justice in their political life" (2005: 77). What makes this a conception of *full* autonomy is that, in accepting and applying principles that they have come to affirm, citizens of the well-ordered society are, Rawls claims, at the 'outer limit' of their freedom. Because the political institutions that we grow up under begin to shape our character and self-conception from so early in our life, the mere fact that we are permitted to emigrate does not suffice to render our acceptance of political authority free. But if "over the course of life [we] come freely to accept as the outcome of reflective thought and reasoned judgment, the ideals, principles, and standards that specify our basic rights and liberties, and effectively guide and moderate the political power to which we are subject" then we render ourselves as free as we can possibly be, given that we inevitably grow up in a particular social and political world with concomitant influences on us (*Ibid.*: 222).

This conception of autonomy is, I believe, distinctive, and its strengths and weaknesses are still relatively underexplored. We can further examine its contours by considering how it relates to the aspects of autonomy discussed above.

Let us begin with *authenticity*. Recall that this aspect of autonomy requires that your desires are your own. You fail to be autonomous in this way if you experience your inclinations as an alien imposition, such as if you are in the thralls of an addiction. Alien desires compromise our autonomy as they prevent us from governing our lives in accordance with our deepest commitments. Rawls is concerned about how our desires can experienced as alien, though his focus is on our desire to act justly. Citizens of a stable well-ordered society enjoy autonomy as authenticity with respect to this desire, as when they reflect on their sense of justice, they see themselves as having reasons to bolster and maintain it.[8]

We should note here that this autonomy that citizens of the well-ordered society enjoy goes beyond the more common idea that coercive political institutions can be experienced as alien. Colonised nations, groups living under military occupation, and secessionists have all couched their desire for independence in terms of a claim to political autonomy. But here their lack of autonomy consists in their living under a political authority that is misaligned with their judgments about who they ought to be governed by: the coercion from their political institutions is alien in the straightforward sense that they reject it.[9] Rawls's political conception of autonomy goes deeper than this. When citizens accept the authority of their political institutions, their autonomy may still be threatened if this acceptance is not product of reflective thought and judgment in conditions of freedom. And even citizens who reject those institutions don't necessarily have their autonomy threatened, provided their organising principles are ones that would be accepted by the citizens of Rawls's well-ordered society.

A challenge that is sometimes raised to the hierarchical analysis of autonomy as authenticity is that an agent could simply bring their first and second-order desires into harmony by rejecting the latter. If the addict lacks autonomy because of his second-order desire not to be addicted, then he can just as easily become autonomous by affirming his addiction as by taking steps to rid himself of it as he can by overcoming his addiction. We might wonder if a similar challenge could be raised against Rawls's view. Why can't a citizen simply render themselves fully autonomous by endorsing the political institutions they've grown up under, whatever they happen to be? Wouldn't she then be affirming the role that her upbringing played in the development of her character and self-conception, leaving her at the outer limit of her freedom?

To see Rawls's answer to this, we need to see how his conception of autonomy goes beyond merely affirming the influences on our desire to act justly.[10] First, he holds that

to be fully autonomous our principles of political justice must be *liberal*, they must be ones that allow us and others to enjoy the protection of rights and liberties. Rawls defines a liberal conception of justice as one that specifies a set of rights, liberties, and opportunities, gives them a special priority over demands to promote the general good, and assures all citizens adequate all-purpose means to make use of their liberties and opportunities (2005: 6). Second, he holds that to be fully autonomous our principles must be ones that would be chosen in the original position. Drawing on the Kantian interpretation, we can say that the principles that would be chosen in the original position are those we would give ourselves when fairly represented as free and equal persons. Third, our principles must be appropriately public. They must satisfy a full publicity condition, meaning that their justification must be fully available to all citizens (*Ibid.*: 66). And in line with his ideal of public reason, major political decisions in our society must be settled by reasons drawn from a political conception of justice (*Ibid.*: 213). Full autonomy on Rawls's view is therefore a demanding ideal. Though affirming the political institutions that we have grown up and that have influenced our character is necessary for our autonomy, it far from sufficient.

An important question raised by this discussion is what it implies for the autonomy of people here and now who do not (of course) inhabit the heavily idealized world of a society well-ordered by Justice as Fairness. Are we thereby necessarily lacking a degree or component of autonomy? One way of answering 'no' to this question would be to argue that persons who accept and act on the principles of Justice as Fairness—even in an unjust and otherwise non-ideal world—are nonetheless autonomous. After all, their political convictions are ones that they *would* affirm in a stable well-ordered society, and that they would affirm them in those circumstances shows that they are not simply the product of ideology or indoctrination. This answer would seem to cohere with Rawls's discussion the Kantian interpretation, where he holds that acting on the principles that would be chosen in the original position is to express our nature

as free and equal rational beings, regardless of the conditions in which we find ourselves.[11] The case for answering 'no', however, is that we do appear to lack something that citizens of the well-ordered society enjoy. After all, they live in a social and political world that is in accordance with their deepest convictions and are so able to affirm the ways in which they have been shaped by their political institutions. For us this is not possible. But to say that we all lack a component of autonomy in this way need not render the underlying conception of autonomy unrealistic or overly utopian.[12] If Rawls's is right that a stable well-ordered liberal society is practically possible, then the social world in which are all fully autonomous is a possible, if very distant, ideal.

Next, let us consider *procedural independence*. Working out what constitutes a failure of procedural independence is a matter of working out which influences on our desires are consistent with our autonomy. If we have some desire only because we have been coerced or manipulated, then our acting on this desire is not autonomous. In his model of the citizens of the well-ordered society coming to accept his principles, Rawls appeals to a version of procedural independence. Stability for the right reasons requires that the desire to act justly comes about freely, rather than as the result of coercion, indoctrination, or manipulation. In specifying what this requires, he holds that the desires that stabilise his principles must come about solely via the educative effective of growing up in a society well-ordered by them (Rawls, 1999: 401). Since the well-ordered society satisfies a full publicity condition, its citizens are aware of the principles of justice, and they know the arguments that speak in favour of them. Rawls thinks that when citizens of such a society go through an upbringing and education in line with these principles, and see their fellow citizens acting in accordance with them, this will lead them endorse their own desire to act justly. Whatever we make of this argument, Rawls is offering an account of how we should distinguish between education and indoctrination in the political domain.

Finally, we can consider *moral autonomy*. As we saw above, many have found the Kantian idea that the moral law is self-legislated paradoxical. If morality is objective in its content, and it applies to us simply in virtue of our status as rational agents, then how can it also be dependent on whether we choose to self-legislate it? There have been numerous Kantian attempts to resolve this difficulty, but here I want to suggest that Rawls's conception of autonomy points toward a distinctive kind of resolution.

Recall that the challenge arises when we consider the question: what if I do not give myself the moral law? What if I choose to self-legislate some other law, or indeed no law at all? When confronted with these questions, the Kantian can say either that it doesn't really matter whether we give ourselves the moral law: morality is objective, universal, and holds regardless of our attitude to it. Or, they can say that it *does* matter, and so jettison the objective character of morality. But neither of these options seems palatable. The former leaves us without any remaining sense of the moral law as self-legislated, and the latter leaves us with a wholly subjective conception of morality.

But let us suppose that it is not us asking this question, but a citizen of Rawls's well-ordered society. If the stability argument succeeds, these citizens do choose to give themselves the principles of justice. As we have seen, what the stability argument aims to show is that all citizens of the well-ordered society would see themselves as having sufficient reason to affirm the principles of justice operative in their society, and to maintain their sense of justice that is based on these principles. This argument of course depends on various conjectures about the psychological development of citizens, and about the nature of the moral and religious doctrines that are likely persist in in a well-ordered society. But let us assume here that it succeeds. If it does, then there is a straightforward sense in which, for the citizens of the well-ordered society, the principles of political justice are self-legislated. And if we take the normative force of those principles to depend not on whether they are self-legislated by us, but on whether they are self-legislated by the citizens of a well-ordered society,

then we can render the objectivity of political morality consistent with its self-legislation.[13]

This might seem quite insufficient, however. Once again, we are not citizens of a well-ordered society, so it seems the principles of justice are not self-legislated for us at all. If moral autonomy requires us to live under principles that we give ourselves, then it appears we still lack it. But I think here again autonomy is best thought of as a (distant) ideal. If a well-ordered society is one in which every citizen can enjoy the autonomy that stems from living under principles of justice that are self-legislated, then this is part of what makes it an attractive ideal at which to aim. The fact that we do not enjoy that kind of autonomy in the world as it is does not speak against its value, or against the value of Rawls's principles that comes from their being uniquely able to realize it.

§5. Objections

We have now seen how a conception of autonomy might continue to pay a major role in the argument of Rawls's political liberalism, and how this conception speaks to the various aspects of self-government that arise in broader discussions of autonomy. However, as we saw above, many readers of *Political Liberalism* have been sceptical that there is any place for autonomy in the argument for political liberalism. There are two ways in which this sceptical thought might be made out, and I will now discuss each of them in turn.

To begin with the first, the central aim of political liberalism to be compatible with the range of views that the reasonable citizens of a liberal society can be expected to hold. Rawls therefore puts his view forward as a 'freestanding' conception, one that does not depend on any doctrine that might be controversial among reasonable citizens (2005: 10). But conceptions of autonomy appear to be precisely the kind of thing that Rawls thought would be controversial among reasonable citizens. Given that reasonable citizens hold a variety of moral and religious doctrines, some of which may

reject the value of autonomy entirely, how could political liberalism depend for its justification on a conception of autonomy (Quong, 2013: 270–271)? That would seem to leave the view inconsistent with its starting motivations.

This is an objection that shares a common structure with many of the central criticisms of political liberalism. Since the theory was first developed, critics have been pointing out that reasonable citizens are likely to have disagreements that are more extensive than Rawls supposed, and that this causes serious problems for the view.[14] In response, the first point to note is that in political liberalism 'reasonable citizen' is a term of art that picks out an idealized consistency. Different defenders of political liberalism have taken it to refer to different idealized consistencies. My preferred reading, which I think fits best with Rawls's text, is that reasonable citizens are those who have grown up in a well-ordered liberal society (Quong, 2011: chapter 5; Weithman, 2011). As has been pointed out elsewhere, this reading makes the political liberal search for stability and the aim of finding principles acceptable to reasonable citizens one and the same thing (Mulhall and Swift, 1992: 186). If we take this reading, then the objection that autonomy cannot play a role in the justificatory apparatus of the theory because reasonable citizens would reject seems to be putting the cart before the horse. The appeal to autonomy is not something that must satisfy some prior requirement of acceptability to reasonable citizens, it is what explains the force of that requirement in the first place.

This reply might be thought to not go far enough. What if the citizens of the well-ordered society would not accept the conception of autonomy set out above? Political liberalism might be thought to be incoherent or self-defeating if it requires acceptability to citizens who reject its foundational arguments. The right response to this worry is to think about what citizens of the well-ordered society could come to accept. Here what is important to note is that the conception of autonomy set out in the previous section is a distinctly political value. It speaks only to the question of how

the acceptance of political authority and principles of political justice can be autonomous, without making any claims about other domains of conduct. In principle, then, it is compatible with a range of different comprehensive world views that the citizens of a well-ordered society might hold. Some may still think it is too optimistic to suppose that this conception of autonomy could be the object of stable agreement among reasonable citizens. But to these critics it must be pointed out that Rawls's more general views about what reasonable citizens could accept are already highly ambitious. They include, among other things, the acceptance of policies such as the public financing of elections and 'society as an employer of last resort' (Rawls, 2005: lvi–lvii). Surely even if agreement on a political conception of full autonomy seems ambitious, it is not considerably more so than agreement on these policies.[15]

A second objection to the role of autonomy in political liberalism protests that citizens of Rawls's well-ordered society would not be autonomous anyway. Rawls admits that there is likely to be reasonable disagreement about justice in any well-ordered society, such that the best we can expect is agreement on a family of liberal conceptions (*Ibid.*: 163). Given this admission, the model of the well-ordered society becomes one in which citizens accept different conceptions of justice from within this family of liberal views, each for different reasons. This means that it is not exactly true that each citizen accepts the conception of justice that was operative during her upbringing. She may endorse one member of the family of liberal conceptions of justice, having been raised in a society that implements another. How could such a citizen be autonomous in the sense of wholeheartedly accepting the social and political forces that have acted on her through her education and upbringing?

This challenge can be answered by considering more fully what those who endorse one of the family of liberal conceptions accept. In addition to accepting a conception of justice that prioritises a set of basic rights and liberties over the general good and guarantees adequate all-purpose means for citizens to make use of their rights,

liberties and opportunities, Rawls also supposes that citizens accept a particular view about how their disagreements about justice ought to be settled. Reasonable citizens accept that when they cannot reach a consensus about issues that are disputed by the different liberal conceptions, they should settle the matter by appeal to majority rule (*Ibid.*: 393). Given this, a citizen who rejects the conception of justice that is operative in her society in favour of another member of the family of liberal conceptions is not alienated from her social and political institutions in a way that impacts her autonomy. After all, she accepts that this disagreement about justice has been dealt with in the right way, and so she endorses the decision-making procedure that led to her society being governed in the way that it is.

§6. Conclusion

Many readers of Rawls's work hold that while he was committed to a Kantian conception of autonomy in *Theory*, by the time of *Political Liberalism* he had left behind his autonomy-based commitments. The goal uncovering principles that could be the object of a consensus among reasonable citizens is thought to force him to jettison autonomy from the justificatory apparatus of the view.

I have argued here in favour of a different picture. Rawls's conception of autonomy presents an ideal in which all citizens freely and reflectively endorse the social and political institutions they have grown up under. A great strength of this conception is that it acknowledges the threat to our autonomy that stems from the ways in which we are deeply shaped by the environment in which our upbringing takes place. Rawls's view acknowledges this threat and aims to show that it is nonetheless possible for our acceptance of political authority to be fully autonomous. It is this aim that explains his commitment to stability, and thus his development of a distinctly political liberalism.

Bibliography

Billingham, P. & Taylor, A. "A Framework for Analyzing Public Reason Theories," *European Journal of Political Theory*, Forthcoming.

Caney, S. "Liberal Legitimacy, Reasonable Disagreement and Justice," *Critical Review of International Social and Political Philosophy* 1 (1998): 19–36.

Clayton, M. *Justice and Legitimacy in Upbringing* (Oxford: Oxford University Press, 2006).

Cohen, J. "Reflections on Rousseau: Autonomy and Democracy," *Philosophy & Public Affairs* 15 (1986): 275–297.

Darwall, S. "A Defense of the Kantian Interpretation," *Ethics* 86 (1976): 164–170.

Dreben, B. "On Rawls and Political Liberalism," in Freeman, S. (ed.), *The Cambridge Companion to Rawls* (New York, NY: Cambridge University Press, 2003).

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

———"Acting Freely," *Nous* 9 (1970): 367–383.

Enoch, D. "Hypothetical Consent and the Value(s) of Autonomy," *Ethics* 128 (2017): 6–36.

Frankfurt, H. "Freedom of the Will and the Conception of a Person," Journal of Philosophy 68 (1971): 5–20.

Freeman, S. "Congruence and the Good of Justice," in Freeman, S. (ed.), *The Cambridge Companion to Rawls* (New York, NY: Cambridge University Press, 2003).

Kleingeld, P. & Willaschek, M. "Autonomy Without Paradox: Kant, Self-Legislation and the Moral Law," *Philosophers' Imprint* 19 (2019): 1–18.

Krasnoff, L. "Consensus, Stability, and Normativity in Rawls's *Political Liberalism*," *Journal of Philosophy* 95 (1998): 269–292.

Mulhall, S. & Swift, A. *Liberals and Communitarians* (Oxford: Blackwell, 1992).

Quong, J. "On the Idea of Public Reason," in Mandle, J. & Reidy, D.A. (eds.), *A Companion to Rawls* (Oxford: Wiley Blackwell, 2013).

———*Liberalism Without Perfection* (New York, NY: Oxford University Press, 2011).

Rawls, J. *Political Liberalism: Expanded Edition* (New York, NY: Columbia University Press, 2005)

——— *Justice as Fairness: A Restatement* (Cambridge, MA: The Belknap Press of Harvard University Press, 2001).

———*A Theory of Justice: Revised Edition* (Cambridge, MA: The Belknap Press of Harvard University Press, 1999).

Raz, J. *The Morality of Freedom* (New York, NY: Oxford University Press, 1986).

Rostbøll, C.F. "Kantian Autonomy and Political Liberalism," *Social Theory and Practice* 37 (2011): 341–364.
Scanlon, T.M. "A Theory of Freedom of Expression," *Philosophy & Public Affairs* 1 (1972): 204–226.
Stilz, A. *Territorial Sovereignty: A Philosophical Exploration* (New York, NY: Oxford University Press, 2019).
Taylor, A. "Stability, Autonomy, and the Foundations of Political Liberalism," *Law and Philosophy*, Forthcoming.
Waldron, J. *Law and Disagreement* (New York, NY: Oxford University Press, 1999).
Wall, S. "Is Public Justification Self-Defeating?" *American Philosophical Quarterly* 39 (2002): 501–507.
Weithman, P. "Autonomy and Disagreement about Justice in *Political Liberalism*," *Ethics* 128 (2017): 95–122.
———*Why Political Liberalism? On John Rawls's Political Turn* (New York, NY: Oxford University Press, 2011).

[1] For the role of this idea in Rousseau, see (Cohen, 1986).
[2] The origins of which are in (Dworkin, 1970) and (Frankfurt, 1971).
[3] For a recent discussion of this difficulty, see (Kleingeld and Willaschek, 2019).
[4] This is the task that Rawls embarks upon in (1999: part III).
[5] For a valuable reconstruction of this argument, see (Weithman, 2011: chapter VII).
[6] This a phrase of Rawls's. See, e.g., (2001: 4).
[7] An autonomy-based argument for Rawls's stability condition is set out in (Clayton, 2006: 11–19). I also offer a defence of it in (Taylor, Forthcoming); the remainder of this section draws on that defence.
[8] Indeed, in Paul Weithman's recounting of Rawls's stability arguments, he describes the way in which citizens reflectively endorse their sense of justice in explicitly hierarchical terms, writing that they have a "highest-order regulative desire" to act justly (2011: 64).
[9] For an exploration of a conception of political autonomy of this kind, see (Stilz, 2019: part II).
[10] Rawls's comments relating to this conception of autonomy are contained primarily in (2005: 77–78). In reconstructing his view, I draw heavily on (Weithman, 2017: 98–105).
[11] For a discussion of what it might mean to express our nature in different circumstances, see (Weithman, 2011: 198).
[12] For this condition on a conception of autonomy, see (Dworkin, 1988: 7–8).
[13] See also (1999: 450–456) for Rawls's comments on the relationship between autonomy and objectivity.
[14] This is at the root of the so-called asymmetry and self-defeat objections to political liberalism. See (Caney, 1998) and (Waldron, 1999: 149–163) on the asymmetry objection, and (Wall, 2002) on self-defeat. For further discussion, see (Billingham and Taylor, Forthcoming).
[15] Another way to put this point is as follows. In order to not be completely sceptical about Rawls's project, we must accept that some substantive and controversial political conclusions can be the object of agreement in a well-ordered society. And among those who are happy to accept *that*, I am not aware of any good argument for being especially sceptical about agreement on the political conception of autonomy I have outlined here.