Kant, Rawls, and the Possibility of Autonomy

By: Brian Kogelmann, Department of Philosophy, University of Maryland

Email: bkogel89@gmail.com

Acknowledgements: The author would like to thank Mark Timmons for his helpful feedback on earlier drafts of this paper.

Abstract: One feature of John Rawls's well-ordered society in both *A Theory of Justice* (*TJ*) and *Political Liberalism* (*PL*) is that citizens in the well-ordered society, when adhering to the principles of justice governing that society, realize their *full autonomy*. This notion of full autonomy is explicitly Kantian. This constancy, I shall argue, raises problems. Though the model of the well-ordered society presented in *TJ* is arguably consistent with Kant's notion of autonomy, the model of the well-ordered society presented in *PL* is not. The problem is that in the well-ordered society of *PL* people's reasons for complying with the principles of justice are overdetermined in a problematic way. This raises the interesting question of acting from overdetermined motives in Kant's system of ethics. In this paper I argue that regardless of which plausible interpretation of acting from overdetermined motives we adopt, the prospect of citizens realizing their full autonomy in Rawls's *PL* are small. This is a serious defect of the theory.

Keywords: Rawls; Kant; political liberalism; stability; autonomy; overdetermination.

1. Introduction

One feature of John Rawls's well-ordered society in both *A Theory of Justice* (*TJ*) and *Political Liberalism* (*PL*) is that citizens in the well-ordered society, when adhering to the principles of justice governing that society, realize their *full autonomy*. This notion of full autonomy is explicitly Kantian (Rawls 1971: §40; Rawls 1980: 315-317, 320; Rawls 1993: 77-81; Freeman 2007b; Forst 2017; Weithman 2017). Just as moral agents are free within Kant's system of ethics when they act in accordance with the self-legislated moral law motivated solely by duty to that law, so too are they in Rawls's well-ordered society (§2). When citizens act in

accordance with the principles of justice they would self-legislate in a position of impartiality (i.e., the original position), and adhere to these principles out of the proper motive, they act from these principles as just and thus autonomously. As Rawls phrases it: "full autonomy is realized by citizens when they act from principles of justice that specify the fair terms of cooperation they would give to themselves when fairly represented as free and equal persons" (Rawls 1993: 77).

Though many things changed about Rawls's theory in the switch from *TJ* to *PL*, Rawls's insistence that citizens in the well-ordered society realize their full autonomy when acting justly remains constant (e.g., Rawls 1980: 315-317, 320; Rawls 1993: 77-81). This constancy, I shall argue, raises problems. Though the model of the well-ordered society presented in *TJ* is arguably consistent with Kant's conception of autonomy, the model of the well-ordered society presented in *PL* is not.[1] The problem is that in the well-ordered society of *PL* citizens' motives for obeying the principles of justice are overdetermined in a problematic way (§3). This raises the interesting question of acting from overdetermined motives in Kant's system of ethics, and the relationship between acting on the basis of overdetermined motives to Kant's conception of autonomy (§4). I shall argue that regardless of which plausible interpretation of Kant we adopt, there is no way we can guarantee that citizens in the well-ordered society of *PL* will realize their full autonomy as Rawls hoped (§5).

If my argument is correct then this is a serious defect with Rawls's latter theory as articulated in *PL*. The idea of the well-ordered society is supposed to embody "certain general features of any society that it seems one would, on due reflection, wish to live in and want to shape our interests and character" (Rawls 1974: 232-233).[2] That is, there are certain normatively attractive features well-ordered societies possess. One supposed feature is that persons in these societies are autonomous. There is, I think, something quite attractive about this notion of citizens being fully autonomous. This is especially so when we consider that the rules we want citizens to be autonomous with respect to are political ones. Indeed, recall a feature of political rules that Rawls often emphasizes: they are typically coercive, and typically have a pervasive

[1] Weithman (2017) explores a very similar question in a recent paper. The difference, though, is that the current paper explores whether Rawls's Kantian conception of autonomy is consistent with how he secures the overlapping consensus in *PL*, and Weithman's paper explores whether this conception of autonomy is consistent with the fact that Rawls admits reasonable disagreements about justice in the introduction to the paperback version of *PL*. This latter question is also examined in Kogelmann (2017).
[2] For similar remarks see Rawls (1975: 254-255); Rawls (1980: 325).

influence on people's life prospects (e.g., Rawls 1980: 326). These twin features make such institutions inherently suspect. One way of responding to the inherently suspect nature of coercive social and political institutions is to insist that persons be autonomous with respect to them; it must be such that persons would give themselves these rules were they to reflect on the matter, and comply with them from the proper motive. This plausibly mitigates some of the badness associated with the coercive and deeply impactful nature of these institutions, for persons in a sense voluntarily comply with such rules when they act autonomously with respect to them. More generally, there has been a recent resurgence in the secondary literature that tries to revive and emphasize the Kantian aspects of Rawls (e.g., Taylor 2011; Forst 2017; Weithman 2017). There is no secret why this is so – many of the Kantian aspects of Rawls's theorizing are deeply attractive. It would thus be unfortunate if persons in Rawls's final articulation of the well-ordered society failed to be autonomous, *contra* what he insists. This, however, is what I shall argue.

2. Kant on Autonomy

Our goal is to see to what extent Rawls's well-ordered society as presented in *PL* is consistent with Kant's conception of autonomy, as Rawls had hoped. Before doing this we need to know a bit more about Kant's conception of autonomy, as adopted by Rawls. Though fleshing out Kant's conception of autonomy could itself be the topic of an entire essay, I take Kant's conception of autonomy as having two essential features: a *positive component* and a *negative component* (Sensen 2015: 263; Taylor 2011: 64; Rawls 2000: 205-207). As far as the positive component goes, autonomous agents adhere to the moral law they give themselves. The autonomous will has a *self-legislating capacity*, the "will's property of being a law to itself," and, furthermore, autonomous agents employ and act on this capacity (GMM 4:447). This positive component of autonomy is captured in Rawls's system by the idea of the original position. The principles of justice governing the well-ordered society are those that deliberators in the original position would self-legislate in an impartial position (Rawls 1971: 252-253). The positive component of Kant's conception of autonomy is satisfied when persons in the well-ordered society act in accordance with these principles.

Yet acting in accordance with these principles is merely necessary but not sufficient for persons to be autonomous. As far as the negative component of Kant's conception of autonomy goes, autonomous agents have the property of independence from "alien causes" (GMM 4:446). That is, when agents act in accordance with the moral law they self-legislate they do so from the proper motive – the motive of duty. As Rawls explicates it, "a supremely legislative will… cannot itself be dependent on any interest. Kant means that such a will cannot be dependent on interests derived from natural desires, but depends solely on interests taken in the principles of practical reason" (Rawls 2000: 206). This negative component of autonomy is captured in Rawls's system when he stipulates that persons realize their full autonomy when they "in their conduct as citizens not *only* comply with the principles of justice, but they also act *from* these principles as just" (Rawls 1993: 77) (emphasis mine). Even if persons act in accordance with the principles of justice they would give themselves in the original position (thus satisfying the positive component of Kant's conception of autonomy), if they do so out of the wrong kind of motives – say, they do so to gain an advantage over their fellow citizens – then they still do not act autonomously, for they do not satisfy the negative component of Kant's conception of autonomy.

More needs to be said here about this negative aspect. In particular, does satisfying the negative component of autonomy mean that persons do not act from desires when acting autonomously? This is important, for most contemporary theories of action hold that persons *always* act from desires of some kind. Rawls is aware of this, and works to further flesh out the negative component of Kant's conception of autonomy. He writes:

> To this the answer is that a Kantian view does not deny that we act from some desire. What is of moment is the kinds of desires from which we act and how they are ordered; that is, how these desires originate within and are related to the self, and the way their structure and priority are determined by principles of justice connected with the conception of the person we affirm… Given this connection, an effective sense of justice, the desire to act from the principle of justice, is not a desire on the same footing with natural inclinations; it is an executive and regulative highest-order desire to act from certain principles of justice in view of their connection with a conception of the person as free and equal. And that desire is not heteronomous: for whether a desire is heteronomous

> is settled by its mode of origin and the role within the self and by what it is a desire for in this case the desire is to be a certain kind of person specified by the conception of fully autonomous citizens of a well-ordered society (Rawls 1980: 320).

Rawls's response here is as follows. When it comes to Kant's notion of acting from the motive of duty and thus the negative component of Kant's conception of autonomy, Rawls argues that a plausible interpretation of Kant says that individuals – even when acting from the motive of duty alone and, indeed, when acting from *any* motive – are still always acting from desires. What matters, though, is the nature of these desires, as well as their origins. So long as the desire one acts from (e.g., the desire to do what is just) is an executive and regulative highest-order desire, and originated via some reflective process rather than just being a mere natural inclination, then it is reasonable to say that acting on such desires is sufficient to satisfy what Kant means when he talks about acting from the motive of duty alone, or acting from "*pure reason*" (MS 6:213).

Call these the *executive requirement* and the *reflective requirement* respectively. A desire satisfies the executive requirement if it is highest-order and regulative, subordinating other desires to it. And a desire satisfies the reflective requirement if it came about via a reflective process, rather than being a mere natural inclination. If one acts from desires satisfying the executive requirement and the reflective requirement then the negative component of Kant's conception of autonomy is satisfied, in that one does not act from alien causes. We shall follow Rawls here and grant that he is correct on this point.[3] Satisfying the negative component of

[3] Some might hold that acting from a motive that satisfies the executive requirement and reflective requirement is still not sufficient to capture what Kant means by acting from the motive of duty and not heteronomous causes. Consider an example. Suppose I have a desire to relentlessly pursue wealth, and this desire satisfies both the executive and reflective requirements. Intuitively, acting on this desire would not count as acting from the motive of duty. It may be, then, that to avoid such cases we need to also add a third requirement restricting the content of desires in order to capture what it means to act from the motive of duty. On this view, the executive and reflective requirements are thus not jointly sufficient but merely necessary to act autonomously.

Articulating this third requirement would take us too far afield for this paper. If one wants to interpret the executive and reflective requirements as necessary but not jointly sufficient then this is fine. The central argument of the paper still stands, in that the *PL* model violates these necessary conditions. The only change here is that it is now an open question whether persons in the *TJ* model act autonomously, for it may be that they violate the third, yet-to-be-articulated requirement. If true it would still be an important difference, however, that persons in the *TJ* model fail to be autonomous for this reason, whereas persons in the *PL* model fail to be autonomous for violating the executive and reflective requirements (these requirements being satisfied in the *TJ* model), as well as perhaps this new third requirement.

Kant's conception of autonomy thus requires that persons act justly for the right reasons, which means that the desires causing them to act justly are desires satisfying the executive and reflective requirements. For the remainder of the paper, we will be asking whether both the positive and negative elements of Kant's conception of autonomy survive Rawls's political liberal turn. If they do not, then the status of Rawls's political liberalism model as a genuine political ideal is cast into doubt.

3. Full Autonomy, Stability, and the Well-Ordered Society

Before asking whether citizens realize their full autonomy in both the well-ordered society of *TJ* and the well-ordered society of *PL* we need to know more about both models. Specifically, we need to better understand why Rawls thinks both models of the well-ordered society will be stable, in the sense that "inevitable deviations from justice are effectively corrected or held within tolerable bounds by forces within the system" (Rawls 1971: 458). As we shall see, though the way in which Rawls secures the stability of the well-ordered society in *TJ* is arguably consistent with Kant's conception of autonomy, the way in which Rawls secures the stability of the well-ordered society in *PL* seems at least *prima facie* inconsistent with Kant's conception of autonomy.

3.1 Stability and Autonomy in *A Theory of Justice*

Rawls's argument for the stability of the well-ordered society in *TJ* proceeds in two stages. In the first stage, Rawls argues that those growing up in the well-ordered society develop a sense of justice via a three-stage developmental process (Rawls 1963; Rawls 1971: ch. 8). This three-stage developmental process inculcates a sense of justice, which means that among citizens in the well-ordered society there is a "desire to do what is just," such that "no one wishes to advance his interests unfairly to the disadvantage of others" (Rawls 1971: 497). Initially one might think that, since citizens have a desire to do what is just via their sense of justice, when citizens do act in accordance with the principles of justice they would give themselves in the original position (thus satisfying the positive component of Kant's conception of autonomy) they

do not do so from the motive of duty alone.[4] Rather, they are much like Kant's individual who complies with the moral law from the motive of sympathy (GMM 4:398). But such an example, according to Kant, is a case of adhering to duty out of improper motives. The sympathetic individual's actions – though they are *in accordance* with the demands of duty – are not autonomous because they are not done *from the motive* of duty.

Whether this criticism holds depends on whether the sense of justice satisfies the executive and reflective requirements, as described in the prior section. Rawls certainly believes this is the case. He says: "an effective sense of justice, the desire to act from the principles of justice, is not a desire on the same footing with natural inclinations; it is an executive and regulative highest-order desire to act form certain principle of justice in view of their connection with a conception of the person as free and equal" (Rawls 1980: 320). The idea is this. The sense of justice satisfies the executive requirement because it is a moral sentiment, which means that it is a "governing disposition"; this is contrast to natural attitudes, which "need not be so regulative or enduring" (Rawls 1971: 479). And the sense of justice satisfies the reflective requirement because it is consciously adopted given our knowledge of the principles of justice governing our society, where such adoption is independent from "accidental circumstances" and "contingencies" (Rawls 1971: 475). As such, when citizens are driven by their sense of justice to act justly, they do so from the proper motive, thus satisfying the negative component of Kant's conception of autonomy.

This is not the whole of the stability story, however. Though citizens in the well-ordered society have a sense of justice and thus act justly, Rawls was concerned with citizens maintaining their sense of justice. Even if it is the case that Althea has a desire to do what is just, it might be that Althea does not desire to maintain such a desire: "Members of the well-ordered society may resent their own sense of justice because of its costs. Once they realize their society is set up to encourage that sentiment, they may worry that they have been illegitimately indoctrinated" (Weithman 2010: 53).[5] According to Paul Weithman's interpretation, Rawls addresses this problem by arguing that citizens would desire to maintain their sense of justice

[4] For such criticism of Rawls's claim that citizens in the well-ordered society realize their autonomy in a Kantian sense see Johnson (1974). For a response see Darwall (1976).
[5] To use the terminology of game theory, this actually presents a tricky assurance problem, the details of which are examined in Kogelmann and Stich (2016); Kogelmann (2019).

from the standpoint of the *thin theory of the good*, which are basic structural as well as substantive commitments every agent includes in her theory of the good in virtue of being a rational agent. This makes the sense of justice doubly stable: "Not only does [justice as fairness] generate its own supportive attitudes [via the sense of justice], but these attitudes are desirable from the standpoint of rational persons [i.e., from the standpoint of the thin theory of the good] who have them when they assess their situations independently from the constraints of justice" (Rawls 1971: 399). So not only do citizens have a desire to act in accordance with the principles of justice via their sense of justice, but they also have a desire to maintain this sense of justice from the perspective of the thin theory of the good.

It is important to note here that the thin theory of the good which Rawls assumes everyone adopts in the well-ordered society of *TJ* is actually quite thick. As Gerald Gaus explicates it, the thin theory of the good makes controversial assumptions concerning the structure of our life plans, the nature of our sociability, our sentiments towards love and friendship as well as our sincerity towards others, our need to exercise more complex capacities over simpler ones, and, perhaps most controversially, our desire to realize and express our nature as free and equal moral persons (Gaus 2014: 239-240). These controversial features constituting the thin theory of the good shall be important later on in understanding Rawls's transition from *TJ* to *PL*.

Above we wondered whether Rawls's insistence on the sense of justice being the motive behind just action is consistent with Kant's conception of autonomy, where autonomous action is carried out from the motive of duty alone. This all depends, recall, on whether the sense of justice satisfies what we have called the executive and reflective requirements. In *TJ*, Rawls clearly thinks that a theory of the good (which, recall, is always grounded in the thin theory) is highest-order and regulative, for a theory of the good is a rational life plan that "establishes the basic point of view from which all judgments of value relating to a particular person are to be made and finally rendered consistent" (Rawls 1971: 409). Moreover, Rawls holds in *TJ* that those elements constituting the thin theory of the good would be chosen from the perspective of "full deliberative rationality": "In brief, our good is determined by the plan of life that we would adopt with full deliberative rationality if the future were accurately foreseen and adequately realized in the imagination" (Rawls 1971: 421). For these reasons the desire to comply with the

dictates of the thin theory of the good likely satisfies both the executive and reflective requirements. By implication, when citizens comply with the dictates of justice from their desire to adhere to their theory of the good, they satisfy the negative component of Kant's conception of autonomy, along with the positive component of Kant's conception of autonomy as well.

On the *TJ* model of stability, then, citizens have two motives for acting justly: first (*i*) from their sense of justice and second (*ii*) from the thin theory of the good.[6] But even though citizens according to the *TJ* model act justly out of two different motives, if we follow Rawls's interpretation of what it takes to act from the motive of duty, then *both* motives qualify as acting from the motive of duty, as both relevant desires satisfy the executive and reflective requirements. We conclude that, since the principles of justice governing the well-ordered society are those principles citizens would self-legislate in a position of impartiality (i.e., the original position), citizens in the well-ordered society of *TJ* satisfy both the positive and negative components of Kant's conception of autonomy. They thus act autonomously.

3.2 Stability and Autonomy in *Political Liberalism*

[6] More accurately, it is the sense of justice that drives one to act justly, and the thin theory of the good that tells one to maintain one's sense of justice (later on in *PL*, it will be one's comprehensive doctrine that tells one to maintain one's sense of justice). The thin theory of the good is thus a contributing reason to one's compliance with justice, but does not cause one to act justly in the same way that the sense of justice does. Given this more careful statement, one might wonder whether this should actually be treated as a case of acting from overdetermined motives. I think that it should, and the best way to see why is to think about an analogous case using Kant's examples.

We know from Kant that the person who complies with the moral law out of sympathy has no moral worth and does not act autonomously. Suppose that an individual, call her Bertha, complies with the moral law out of duty. Further suppose that when it comes to *why* Bertha maintains her sense of duty, she does so out of sympathy for her fellow human beings – being sympathetic to their plight requires her to act from the motive of duty, which is the primary driver of her compliance to the moral law. Here the question is: do Bertha's actions have moral worth and should they be considered autonomous in the Kantian framework? The answer here is far from obvious. But certainly, it would be inappropriate to judge Bertha's behavior assuming that duty *and only* duty drive her behavior. This would fail to make her case distinct from that of Cassidy, who complies with the moral law from the motive of duty, and does not require further motivation to maintain her sense of duty. Duty alone is sufficient.

Given this, I think the most obvious approach here would be to examine Bertha's behavior from the perspective of acting from overdetermined motives, even though it is technically only the motive of duty that causes Bertha's compliance to the moral law. Likewise, I think it is quite natural to examine persons in Rawls's well-ordered society as acting from overdetermined motives, even though it is not technically correct to say that both motives cause compliance with justice – rather, both motives *contribute* in some looser sense to one's compliance with justice.

Rawls says that the shift from *TJ* to *PL* has to do with dissatisfaction in his account of the stability of the well-ordered society as presented in *TJ*. Rawls began realizing that this account of stability fails because it relies on the controversial thin theory of the good that could not withstand scrutiny in a liberal society – it is simply not true that all citizens in the well-ordered society would adopt theories of the good containing those substantive and structural commitments articulated in the thin theory. If people adopt conceptions of the good not grounded in the thin theory then, though the sense of justice might be inculcated, it will not necessarily be reinforced as Rawls argued it would. Though citizens might develop a sense of justice they won't necessarily desire to maintain this sense of justice upon reflection. Because the account of stability in *TJ* relied on the controversial thin theory of the good Rawls thought that "the text regards justice as fairness and utilitarianism as comprehensive, or partially comprehensive doctrines" (Rawls 1993: xvi).[7] Abandoning the thesis that all conceptions of the good are grounded in the thin theory, Rawls sought to offer a political conception of justice which could be endorsed by a plurality of comprehensive doctrines rather than just those grounded in the thin theory. Instead of endorsing justice as fairness from within the thin-but-controversial theory of the good, members of a pluralistic society could endorse justice as fairness from within their own comprehensive moral and philosophical doctrines, which may or may not be grounded in the thin theory.

This revision led to a new account of the stability of the well-ordered society. In *TJ* we justify the principles of justice to the citizens of the well-ordered society by appeal to the original position thought experiment: citizens, in the original position, would self-legislate Rawls's two principles of justice. From there those with a sense of justice would desire to act in accordance with these principles, and those with a sense of justice would desire to maintain their sense of justice due to their commitment to the thin theory of the good. In *PL* things are a little more complicated. As before, we begin by appealing to the original position. We first ask whether, from the perspective of deliberators in the original position, citizens would self-legislate Rawls's two principles of justice. This is called *pro tanto* justification. But now, because we need to make sure the principles of justice are justified to several different comprehensive doctrines

[7] Though, as Barry (1995: 887) and Freeman (2007a: 182) remark, it is not quite right to say that justice as fairness *itself* was presented as a comprehensive doctrine. Rather, in order for justice as fairness to be stable it must be *interpreted* as a comprehensive doctrine.

rather than just those grounded in the thin theory, we add a new, second stage of justification: "In this case, the citizen accepts a political conception [of justice] and fills out its justification by embedding it in some way into the citizen's comprehensive doctrine as either true or reasonable, depending on what the doctrine allows" (Rawls 1993: 386). That is, citizens must ask from the perspective of their own unique comprehensive doctrines whether they still have reason to act in accordance with the political conception of justice and maintain their sense of justice.

This second stage of justification in *PL* may raise problems for Rawls's thesis that citizens in the well-ordered society realize their full autonomy when acting justly. In *TJ*, citizens comply with the principles of justice because they have a desire to act justly and a desire to act in accordance with the thin theory of the good. In *PL* citizens comply with the principles of justice because they have a desire to act justly and a desire to act in accordance with their comprehensive doctrine. So, if Althea is a Christian in Rawls's well-ordered society of *PL*, then when Althea adheres to the principles of justice she does so because she has a desire to act justly given her sense of justice, and also because she has a desire to serve her God and believes that adhering to the two principles of justice is the best way of doing so. Like the *TJ* model, Althea's motives for acting justly are overdetermined, but the nature of the second motive is different in *PL*. In *TJ,* citizens act justly because (*a*) they have a desire to do what is just via their sense of justice, and (*b*) they have a desire to comply with the thin theory of the good. In *PL*, citizens act justly because (*a*) they have a desire to do what is just via their sense of justice, and (*c*) they have a desire to act in accordance with their comprehensive doctrines.

The worry here has to do with the stringent requirements Kant places on the negative component of autonomy. If a person does the right thing – say, when a shopkeeper gives correct change – from some motive other than the motive of duty – say, because they want to be a trustworthy shop in the community – then such an action is not autonomous. We cashed this negative component of Kant's conception of autonomy out in terms of two requirements on the desires one acts from: to satisfy the negative component, the desires one acts from must satisfy the executive and reflective requirements. We saw that the sense of justice and the motive to adhere to the thin theory of the good do indeed satisfy these two requirements in *TJ*, meaning that citizens in Rawls's well-ordered society of *TJ* satisfy the negative (as well as the positive) component of Kant's conception of autonomy. Since the sense of justice is one of the reasons

citizens act justly in the well-ordered society of *PL* (carried over from *TJ*), our main question is thus whether the desire to comply with one's comprehensive doctrine satisfies the executive and reflective requirements, as the desire to comply with the thin theory of the good in *TJ* does.

To answer this question more needs to be said about comprehensive doctrines, and why the desire to comply with one's comprehensive doctrine may fail to satisfy either the executive or reflective requirements. First, we must note that a desire to comply with one's comprehensive doctrine as Rawls defines these doctrines can indeed satisfy both the executive and reflective requirements. Consider an example of a possible comprehensive doctrine in the well-ordered society of *PL* given by Samuel Freeman: "Liberal Thomists, then (to take one example), will affirm the principles of justice for their own specific reasons. They are seen as among the natural law preordained by God and knowable by the natural light of our reason" (Freeman 2007a: 169). Intuitively such an individual's desire to adhere to natural law is likely supremely regulative, meaning that such a desire satisfies the executive requirement. And such a desire also seems to have come about from a reflective process – for God's will is knowable from "the natural light of our reason" – meaning that the desire to adhere to natural law also satisfies the reflective requirement. So this is an example of a desire to comply with one's comprehensive doctrine satisfying the negative component of Kant's conception of autonomy. The liberal Thomist, when she acts justly, would thus be autonomous.

But even though an individual's desire to adhere to their comprehensive doctrine *can* satisfy the executive and reflective requirements, there is nothing about how Rawls defines reasonable persons and reasonable comprehensive doctrines *requiring* that a desire to adhere to one's comprehensive doctrine *must* satisfy the executive and reflective requirements, as seemed to be the case with the thin theory of the good in *TJ*. When it comes to comprehensive doctrines, all we are told is that reasonable comprehensive doctrines (*i*) contain a theoretical component, (*ii*) contain a practical component, and (*iii*) belong to a tradition of thought (Rawls 1993: 59). And when it comes to reasonable persons, such persons are defined as "ready to propose principles and standards as fair terms of cooperation and to abide by them willingly, given the assurance that others will likewise do so" (Rawls 1993: 49). Clearly, one could be a reasonable person according to the definition just offered, with a desire to adhere to a comprehensive

doctrine as just defined, where such a desire does not satisfy the executive and reflective requirements.

Indeed, suppose that Althea, our Christian, has a desire to adhere to Protestant doctrine. Suppose that this desire is supremely regulative, meaning that it satisfies the executive requirement. But also suppose that this desire is a product of habituation. Althea never thought deeply about her Protestantism, but rather simply followed the teaching of her parents and community. As a result, such a desire does not satisfy the reflective requirement – for it seems to be a natural inclination – *even though* Althea as described counts as a reasonable person with a reasonable comprehensive doctrine. Since it is possible to be a reasonable person with a reasonable comprehensive doctrine in Rawls's well-ordered society of *PL* while also failing to satisfy the reflective requirement, it follows that citizens in Rawls's well-ordered society of *PL* may fail to satisfy the negative component of Kant's conception of autonomy. When complying with the principles of justice, such citizens are not autonomous.

But recall that the desire to comply with one's comprehensive doctrine is but one motive for acting justly in the well-ordered society of *PL*. Citizens also have a desire to do what is just via their sense of justice which *does* satisfy the executive and reflective requirements. In other words, citizens in the well-ordered society of *PL* act justly from overdetermined motives (as they do in *TJ*), where *one* of these motives satisfies the negative component of Kant's conception of autonomy and where the other motive may (but not necessarily) fail to satisfy the negative component of Kant's conception of autonomy. Thus, whether Althea acts autonomously when she complies with the governing conception of justice in *PL* depends on the status of acting from overdetermined motives in Kant's system of ethics.[8] If we can find an interpretation of Kant where acting from overdetermined motives of the kind just described still results in autonomous action then perhaps Rawls's claim that citizens like Althea realize their full autonomy when

[8] Another possible response is to say that acting from one non-heteronomous motive is sufficient to be autonomous, regardless the nature of the other motives one acts from. This would be a Kantian view, but not Kant's view. Rawls could of course make this move to solve his problem. Yet it would be strange and rather arbitrary given his close adherence to Kant's negative understanding of autonomy up to this point. In order for this kind of revision to appear justified and not *ad hoc*, we would thus need to see some kind of principled reason for revising Kant's account of autonomy in this way, especially given the close reliance on Kant up until now.

acting in accordance with the principles of justice stands. This is what we now turn our attention to.

4. Autonomy, Moral Worth, and Overdetermination: Three Interpretations

In examining whether citizens in the well-ordered society of *PL* realize their full autonomy when acting in accordance with the principles of justice we need to know more about autonomy and acting from overdetermined motives in Kant. As mentioned in §2, we have been understanding autonomy in Kant as having two essential features: a positive component and a negative component. As far as the positive component goes, an autonomous will has a self-legislating capacity, and, moreover, autonomous agents act on this capacity, in that an autonomous agent gives herself and then complies with the moral law. As far as the negative component goes, autonomous actions are independent of alien causes – such actions are caused by the motive of duty which, following Rawls, we determined is equivalent to being caused by desires satisfying the executive and reflective requirements. But what happens when an action is caused *both* by the proper motive *and* an alien cause, which, we showed in the previous section, is possible in the well-ordered society of *PL*? Are such actions autonomous?

Relevant here is the literature on acting from overdetermined motives in Kant's system of ethics. One of the striking features is the vanishingly small number of cases in which actions are, according to Kant, autonomous. Famously, Kant gives cases where individuals perform the morally correct action but do so out of the wrong motive. The shopkeeper, for instance, does not "overcharge an inexperienced customer, and where there is a good deal of trade a merchant does not overcharge but keeps a fixed general price for everyone" (GMM 4:397). The reason the shopkeeper does not overcharge, though, is because he is concerned with his reputation and the future of his business. In such a case the shopkeeper fails to act from the proper motive; as a result, his actions do not have moral worth and are thus not autonomous. Perhaps even more striking is the individual who only does the right thing out of sympathy: "there are many souls so sympathetically attuned that, without any other motive of vanity or self-interest they find an inner satisfaction in spreading joy around them and can take delight in the satisfaction of others

so far as it is their own work" (GMM 4:398). Even in this case, though, Kant says that such action "has nevertheless no true moral worth" (GMM 4:398).

What is absent in Kant is a discussion of the relationship between moral worth, autonomy, and actions resulting from overdetermined motives: what happens if the grocer is motivated by duty *and* the profit motive? And what if the individual in the second case is motivated by duty *and* sympathy? The closest Kant comes to discussing acting from overdetermined motives and moral worth is when he says the following:

> Suppose, then, that the mind of this philanthropist were overclouded by his own grief, which extinguished all sympathy with the fate of others, and that while he still had the means to benefit others in distress their troubles did not move him because he had enough to do with his own; and suppose that now, when no longer incited to it by any inclination, he nevertheless tears himself out of this deadly insensibility and does the action without any inclination, simply from duty; then the action first has its genuine moral worth (GMM 4:398).

As Richard Henson notes, "surely the most obvious way of generalizing from this remark yields the doctrine that only when one acts from duty alone – 'without *any* inclination' – does his act have moral worth" (Henson 1979: 45). This is indeed a natural interpretation of what Kant thinks about actions caused by overdetermined motives: that *any* time motives are overdetermined in the manner we are considering, the resulting action has no moral worth and is thus not autonomous, because it is "simply from duty" that must cause morally praiseworthy and thus autonomous actions. It would be a mistake, however, to use this single passage alone to interpret Kant on this important issue.

This lacuna in Kant has led several Kant scholars to develop accounts of when acting from overdetermined motives have moral worth and are thus autonomous. Perhaps the first attempt at doing so was carried out by Henson, who developed two different accounts. The first is called the *fitness report account* (FRA). It says:

> A dutiful act was done from duty and thus has moral worth provided that respect for duty was present and would have sufficed by itself, even though (as it happened) other motives were also present and might themselves have sufficed (Henson 1979: 48).

The FRA essentially gives a counterfactual test. When there are two motives M_1 and M_2 that cause action A, and M_1 is the motive of duty (in our terminology, M_1 is a desire satisfying the executive and reflective requirements), then A has moral worth and is thus autonomous just in case, were M_2 not present, M_1 would be sufficient to still cause A. Henson's second account is called the *battle citation account* (BCA). It goes as follows:

> A dutiful act was done from duty and thus has moral worth only if respect for duty was the sole motive tending in the direction of the dutiful acts (Henson 1979: 48).

On the BCA, when two present motives (M_1 and M_2) support action A, and M_1 is the motive of duty (in our terminology, M_1 is a desire satisfying the executive and reflective requirements), then an action A has moral worth and is thus autonomous just in case M_1 and *only* M_1 causes A. Unlike the FRA, on the BCA the second non-duty motive is present but not efficacious in that it does not actually cause action A. According to the FRA, both the duty motive and non-duty motive can be efficacious (in that they both cause action A) for the resulting action to still have moral worth and thus be autonomous. Clearly, the BCA is a more demanding test than the FRA in that it is more difficult to satisfy.

Barbara Herman does not think the less-demanding FRA is a plausible contender as an interpretation of Kant. Her reasoning is as follows. Before determining whether the FRA is a plausible contender as an interpretation of when acting from overdetermined motives has moral worth we need to know why Kant thinks the actions of the shopkeeper and the sympathetic man have no moral worth and are thus not autonomous in the first place. According to Herman, Kant thinks these actions have no moral worth because they are the result of good fortune – it is by mere luck that giving correct change is in the shopkeeper's interest. And in the case of the man of sympathy, it is by mere luck that (*i*) the man is sympathetic and (*ii*) that sympathy requires what duty requires: such a disposition "is on the same footing with other inclinations, for example, the inclination to honor, which, *if it fortunately lights upon what is in fact in the common interest and in conformity with duty and hence honorable*, deserves praise and encouragement but not esteem" (GMM 4:398) (emphasis mine).

Kant's remarks suggest a more general thesis for determining when actions fail to have moral worth and are thus not autonomous: "Nonmoral motives may well lead to dutiful actions,

and may do this with any degree of regularity desired. The problem is that the dutiful actions [where no dutiful motives are present] *are the product of a fortuitous alignment of motives and circumstances*. People who act according to duty from such motives may nonetheless remain morally indifferent" (Herman 1982: 366) (emphasis mine). The problem with the FRA account is that it is by mere accident the moral and non-moral motives cooperate in the way they do in the particular situation examined: "For the most part the two motives will cooperate to produce the same action only by accident. As circumstances change we may expect the actions the two motives require to be different and, at times, incompatible" (Herman 1982: 367).

Consider an example. Suppose the shopkeeper has a motive to make a profit and a motive to obey the moral law. Suppose the shopkeeper acts on both motives. Then, on the BCA, the shopkeeper's actions do not have moral worth and are thus not autonomous. What does the FRA say about the shopkeeper's actions? Here we need to ask what happens if the non-duty motive is removed. Suppose that, were we to remove the profit motive, the shopkeeper would still obey the moral law from the motive of duty alone. If true, then the shopkeeper's actions do have moral worth (according to the FRA) when he is driven by both motives, because the relevant counterfactual test is met. Herman's point is that in such a case it is merely fortunate that the profit motive happened to line up with what duty required – it *could* have been the case that the profit motive worked *against* what duty requires. If true then even though the FRA designates the shopkeeper's behavior in circumstance C_1 as having moral worth, it is the case that in some circumstance C_2, the shopkeeper – *with the very same mixed motives approved of by the FRA in* C_1 – would *not* do what duty requires. In Herman's words: "But the fact that the moral motive was sufficient by itself in the overdetermined case does not imply that he would perform honest actions when the profit motive clearly indicated that he should *not* act honestly" (Herman 1982: 368). On Herman's view, the FRA is thus not a plausible interpretation of Kant on the moral worth of acting from overdetermined motives, because it fails to take into account Kant's reasons why the cases he says lack moral worth do indeed lack moral worth. The shopkeeper and man of sympathy's actions lack moral worth and are thus not autonomous because they are only in accordance with duty accidentally; yet on the FRA, morally right actions caused by fortuitous circumstances of this kind have moral worth and are thus autonomous.

Herman does, though, alter the FRA in a manner that is consistent with Kant, called the *greater strength fitness-ready account* (GSFRA). The account goes as follows:

> On a greater-strength interpretation of the fitness model, an action can have moral worth only if the moral motive is strong enough to prevail over the other inclinations [in all possible circumstances]—without concern for whether they in fact cooperate or conflict (Herman 1982: 368).

When there are two motives M_1 and M_2 that cause action A, and M_1 is the motive of duty (in our terminology, M_1 is a desire satisfying the executive and reflective requirements), then A has moral worth and is thus autonomous just in case, in different circumstances where M_2 were to actually conflict with M_1 instead of codetermining A, then M_1 would still cause A. In such a case, "the success of the moral motive was not dependent on the accident of circumstances that produced cooperation rather than conflict" (Herman 1982: 368). The BCA and GSFRA are the two plausible ways of understanding the relationship between acting from overdetermined motives, moral worth, and autonomy in Kant. We now examine whether citizens in the well-ordered society of *PL* act autonomously when they comply with the principle of justice according to either the BCA or the GSFRA. If they do not, then citizens in the well-ordered society of *PL* may not realize their full autonomy, *contra* Rawls's claim, and *contra* the aspirations of his theory.

5. Full Autonomy and the Well-Ordered Society of *Political Liberalism*

In the past section we examined two ways of understanding the relationship between the negative component of Kant's conception of autonomy and acting from overdetermined motives. We saw that there are two plausible interpretations detailing when actions overdetermined by the kinds of motives we are interested in are still autonomous nonetheless. On the BCA, when two present motives (M_1 and M_2) support action A, and M_1 is the motive of duty (in our terminology, M_1 is a desire satisfying the executive and reflective requirements), then an action A has moral worth and is thus autonomous just in case M_1 and *only* M_1 causes A. Here, M_2 – a motive that does not satisfy what we have called the executive and reflective requirements – cannot be causally efficacious in determining action A. On the GSFRA, when there are two motives M_1

and M_2 that cause action A, and M_1 is the motive of duty (in our terminology, M_1 is a desire satisfying the executive and reflective requirements), then A has moral worth and is thus autonomous just in case, in different circumstances where M_2 were to actually conflict with M_1 instead of codetermining A, then M_1 would still cause A. Here, M_2 – a motive that does not satisfy what we have called the executive and reflective requirements – can be causally efficacious in determining action A so long as M_1 wins out over M_2 in counterfactual cases where these two motives conflict with one another. We now examine whether citizens in Rawls's well-ordered society of *PL*, when complying with the principles of justice, act autonomously according to either the BCA or GSFRA.

Consider first the BCA, which says that when two present motives (M_1 and M_2) support action A, and M_1 is the motive of duty (in our terminology, M_1 is a desire satisfying the executive and reflective requirements), then an action A has moral worth and is thus autonomous just in case M_1 and *only* M_1 causes A. In other words, Althea acts autonomously in the well-ordered society of *PL* according to the BCA just in case she has two motives present to act justly – recall, her sense of justice which satisfies the executive and reflective requirements and her desire to adhere to her comprehensive doctrine which, by hypothesis, fails to satisfy the reflective requirement – but *only* her sense of justice causes her to act justly, where her motive to comply with her comprehensive doctrine is not causally efficacious in her complying with the principles of justice. We now ask: is this a plausible interpretation as to why Althea complies with the principles of justice in the well-ordered society of *PL*?

I do not think so. To see why, begin by recalling our discussion concerning why Rawls transitioned from *TJ* to *PL* in the first place. In *TJ*, Rawls argues that the well-ordered society will be stable for two reasons: citizens first develop a sense of justice via a three-stage developmental process, and then decide to maintain this sense of justice from the standpoint of the thin theory of the good. This second step is important, for "members of the well-ordered society may resent their own sense of justice because of its costs. Once they realize their society is set up to encourage that sentiment, they may worry that they have been illegitimately indoctrinated" (Weithman 2010: 53). In other words, the motive to act justly grounded in the thin theory of the good is essential for securing the stability of the well-ordered society, for the sense of justice *on its own* cannot guarantee compliance with the governing conception of justice. It is

the sense of justice *along with* the thin theory of the good guaranteeing that citizens act justly. The sense of justice by itself cannot do this.

But once Rawls abandons the thin theory of the good in the political liberal turn this worry is still a genuine worry. How can we be sure citizens desire to maintain their sense of justice (for the sense of justice alone will not guarantee that citizens act justly)? The answer is that each citizen's comprehensive doctrine also tells them to act justly, thus supporting the sense of justice. But this means that when citizens comply with the principles of justice, such action is caused by *both* their sense of justice – satisfying the executive and reflective requirements – and their desire to comply with their comprehensive doctrines – which, we saw, may fail to satisfy the negative component of Kant's conception of autonomy. Though Althea's motive M_1 (the sense of justice) tells her to act justly, Rawls is worried that she might resent this fact, suggesting that M_1 is not sufficient to guarantee that Althea complies with the principles of justice. As such, M_2 (Althea's desire to comply with her comprehensive doctrine) is playing some kind of efficacious role in Althea's acting justly. But this just means that Althea does not act autonomously according to the BCA. For the BCA says that though M_1 and M_2 are both present, M_2 may play no efficacious role in causing action A – motive M_1 must do all the work. But we have just seen that this cannot be an accurate interpretation of Althea's actions in the well-ordered society of *PL*. On Rawls's model of the stability of the well-ordered society of *PL*, M_2 plays an efficacious role in Althea's acting justly. As such, Althea does not act autonomously according to the BCA.

Though the BCA does not deem Althea's behavior to be autonomous, perhaps the GSFRA does. The GSFRA says that when there are two motives M_1 and M_2 that cause action A, and M_1 is the motive of duty (in our terminology, M_1 is a desire satisfying the executive and reflective requirements), then A has moral worth and is thus autonomous just in case, in different circumstances where M_2 were to actually conflict with M_1 instead of codetermining A, then M_1 would still cause A. In other words, Althea acts autonomously in the well-ordered society of *PL* according to the GSFRA just in case she has two motives that jointly cause her to act justly – recall, her sense of justice which satisfies the executive and reflective requirements and her desire to adhere to her comprehensive doctrine which, by hypothesis, fails to satisfy the reflective requirement – but, were these two motives to actually conflict, then Althea would

continue to act justly out of her sense of justice rather than comply with the dictates of her comprehensive doctrine. So Althea's acting justly from codetermined M_1 and M_2 is autonomous if, in cases where the sense of justice conflicts with the demands of her God telling her to act unjustly, she acts justly nonetheless.

In determining whether Althea's compliance with the principles of justice in the well-ordered society of *PL* satisfies the GSFRA there are two relevant questions to ask. First, we must ask what Rawls thought someone like Althea would do when her comprehensive doctrine and the political conception of justice conflict. Second, we must ask whether what Rawls says is plausible. It could be that, though Rawls believes Althea would act counter to her comprehensive doctrine when her comprehensive doctrine conflicts with the political conception of justice (meaning that Althea's compliance with the principles of justice satisfies the GSFRA), Rawls is mistaken about this.

First: what does Rawls say Althea will do when Althea's comprehensive doctrine conflicts with the political conception of justice? Note here that Rawls does believe this is possible in the well-ordered society of *PL*. He notes that the political conception of justice must only "not conflict too sharply" with citizens' comprehensive doctrines, suggesting that some conflicts are indeed possible (Rawls 1993: 40). But what happens when there is conflict? Writes Rawls:

> The virtues of political cooperation that make a constitutional regime possible are, then, very great virtues. I mean for example, the virtues of tolerance, being ready to meet others halfway, and the virtue of reasonableness and the sense of fairness. When these virtues are widespread in society and sustain its political conception of justice, they constitute a very great public good, part of society's political capital. Thus, the values that conflict with the political conception of justice and its sustaining virtues may be normally outweighed because they come into conflict with the very same conditions that make fair social cooperation possible on a footing of mutual respect (Rawls 1993: 157).

Rawls's answer, then, is that the goods sustained by compliance to the political conception of justice are so important that, when there is conflict between the demands of justice and the demands of one's comprehensive doctrine, citizens will side with the political conception of

justice rather than their comprehensive doctrines. So, when there is conflict between what the political conception of justice tells Althea to do and what her Protestant doctrine tells Althea to do, Althea – because she recognizes just how important adherence to the political conception of justice is – will adhere to the political conception instead of Protestant doctrine. From this we can generalize: there are two motives M_1 and M_2 that cause Althea to adhere to the political conception of justice, and M_1 is the motive of duty. Althea's adherence to the political conception of justice is thus autonomous because, according to Rawls, in different circumstances where M_2 were to actually conflict with M_1 instead of codetermining A, then M_1 would still cause Althea to act justly, for the values in the political conception are "very great virtues."

Is Rawls's argument convincing here? Some have argued that it is not.[9] The first thing to point out is that telling us that the political values are very great values and that they secure important features of social cooperation tells us almost nothing about whether these values and their underlying goods will actually be supplied. Indeed, Rawls's likening of the political values to public goods is quite telling. On standard analyses of public goods, the provision of a public good – say, the provision of a stable constitutional order – is an *n*-person prisoner's dilemma: everyone most prefers that the good is provided without having to actually participate in provision themselves (Hardin 1971). The result is that the good is not provided. Rawls explains the reasoning here: "whatever one man does his action will not significantly affect the amount produced. He regards the collective action of others as already given one way or the other. If the public good is produced his enjoyment is not decreased by his not making a contribution" (Rawls 1971: 267).

Returning to the well-ordered society of *PL*, saying that a stable constitutional order is a public good thus suggests that Althea most prefers to not provide this good – say, by adhering to Protestant doctrine when it conflicts with the political conception of justice – while everyone else provides it by adhering to the political conception. If this is an accurate characterization of Althea's motives, then the GSFRA does not describe her as acting autonomously, for Althea would indeed comply with her comprehensive doctrine over the political conception were the two to conflict. Rawls could get around this problem by arguing that the provision of a just constitutional order is not a public good. But what would he have to assume to make this true?

[9] See here Freyenhagen (2011); Georgieva (2015).

For this to be true Rawls would need to assume that citizens like Althea are *so* committed to the political conception of justice that they would rather comply with the political conception over their comprehensive doctrines when the two conflict, *even when* everyone else adheres to their comprehensive doctrines over the political conception. Then, and only then, could Rawls get around the above objection. Then, and only then, could Rawls say that citizens in the well-ordered society of *PL* satisfy the account of autonomy detailed in the GSFRA.

But what would assuming this entail? It would entail assuming that everyone in society would hold the political conception of justice to be more important than their comprehensive doctrines.[10] Recall, though, from §3.2, why Rawls made the political liberal turn in the first place. In the original model of the well-ordered society of *TJ* Rawls assumed that everyone endorsed the thin theory of the good, which he later realized was incompatible with a liberal society: many people will adopt different values, and at the very least many people will order their values differently. This response to the above, objection, though, assumes something very similar. Not only does this response assume that all members of society endorse the values implicit in the political conception of justice (something the well-ordered society of *PL* must always assume), but it must also assume that citizens order these values in the exact same way: all citizens must place the values of the political conception of justice over the values implicit in their comprehensive doctrines. This, it might be thought, is inconsistent with taking seriously the sort of diversity we find in contemporary liberal societies that motivated Rawls's political liberal turn in the first place.

Not only would this assumption be at odds with Rawls's stated project, it would also be at odds with Rawls's text. In characterizing the kinds of citizens we can expect to find in the well-ordered society Rawls tells us they "are ready to propose principles and standards as fair terms of cooperation and to abide by them willingly, *given the assurance that others will likewise do so*" (Rawls 1993: 49) (emphasis mine). This implies that values simply cannot be structured in the sort of way Rawls must assume they are structured in order for the GSFRA to deem Althea's actions as autonomous: it is simply false that M_1 will always take such precedence that Althea will participate in the provision of public goods even when M_2 says not to. As such, citizens in the well-ordered society of *PL* do not act autonomously according to the GSFRA.

[10] Some followers of Rawls are explicit about this assumption. See, for instance, Larmore (1990: 350).

Since citizens in the well-ordered society of *PL* fail to be autonomous according to both the BCA and GSFRA, it follows that citizens in the well-ordered society of *PL* may not satisfy the negative component of Kantian autonomy. This means that they are not realizing their full autonomy, in contradiction to what Rawls claims.

6. Conclusion

This paper examined the Kantian roots of Rawls's theory of justice, particularly the final version of his theory of justice as presented in *Political Liberalism*. More specifically, we asked to what extent Rawls is correct in saying that citizens in the well-ordered society, as presented in *Political Liberalism*, realize their full autonomy understood in a Kantian sense when they act in accordance with the principles of justice. We concluded that Rawls's claim cannot be sustained. Given the model of society presented in *Political Liberalism*, citizens adhering to the principles of justice are acting from overdetermined motives. Though acting from overdetermined motives is not *as such* a problem for Kant's conception of autonomy, we found that all plausible interpretations of acting autonomously from overdetermined motives in Kant failed to describe the behavior of some citizens in Rawls's well-ordered society. This means that such citizens do not act autonomously, *contra* what Rawls wishes to claim.

Works Cited

Barry, Brian. 1995. "John Rawls and the Search for Stability." *Ethics* 105: 874-915.

Darwall, Stephen L. 1976. "A Defense of the Kantian Interpretation." *Ethics* 86: 164-170.

Forst, Rainer. 2017. "Political Liberalism: A Kantian View." *Ethics* 128: 123-144.

Freeman, Samuel. 2007a. *Justice and the Social Contract*. Oxford: Oxford University Press.

Freeman, Samuel. 2007b. "The Burdens of Justification." *Politics, Philosophy, & Economics* 6: 5-43.

Freyenhagen, Fabian. 2011. “Taking Reasonable Pluralism Seriously.” *Politics, Philosophy, & Economics* 10: 323-342.

Gaus, Gerald. 2014. “The Turn to a Political Liberalism.” In *A Companion to Rawls,* edited by Jon Mandle and David A. Reidy: 235-250. Chichester: John Wiley & Sons.

Georgieva, Mihaela. 2015. “Stability and Congruence in Political Liberalism.” *Political Studies* 63: 481-494.

Hardin, Russell. 1971. “Collective Action as an Agreeable *n*-Prisoners’ Dilemma.” *Systems Research and Behavioral Science* 16: 472-481.

Henson, Richard G. 1979. “What Kant Might Have Said.” *Philosophical Review* 88: 39-54.

Herman, Barbara. 1981. “On the Value of Acting from the Motive of Duty.” *Philosophical Review* 90: 359-382.

Johnson, Oliver A. 1974. “The Kantian Interpretation.” *Ethics* 85: 58-66.

Kant, Immanuel. *Groundwork of the Metaphysics of Morals* (GMM). In *Practical Philosophy: The Cambridge Edition of the Works of Immanuel Kant,* edited and translated by Mary J. Gregor: 37-108. Cambridge: Cambridge University Press.

Kant, Immanuel. *The Metaphysics of Morals* (MS). In *Practical Philosophy: The Cambridge Edition of the Works of Immanuel Kant,* edited and translated by Mary J. Gregor: 353-404. Cambridge: Cambridge University Press.

Kogelmann, Brian. 2017. “Justice, Diversity, and the Well-Ordered Society.” *The Philosophical Quarterly* 67: 663-684.

Kogelmann, Brian. 2019. “Public Reason’s Chaos Theorem.” *Episteme* 16: 200-219.

Kogelmann, Brian and Stephen G.W. Stich. 2016. “When Public Reason Fails Us: Convergence Discourse as Blood Oath.” *American Political Science Review* 110: 717-730.

Larmore, Charles. 1990. “Political Liberalism.” *Political Theory* 18: 339-360.

Rawls, John. 1963/1999. “The Sense of Justice.” In *Collected Papers*, edited by Samuel Freeman: 96-116. Cambridge: Harvard University Press.

Rawls, John. 1971. *A Theory of Justice*. Cambridge: Cambridge University Press.

Rawls, John. 1974/1999. “Reply to Alexander and Musgrave.” In *Collected Papers*, edited by Samuel Freeman: 232-253. Cambridge: Harvard University Pres .

Rawls, John. 1975/1999. “A Kantian Conception of Equality.” In *John Rawls: Collected Papers*, edited by Samuel Freeman: 254-266. Cambridge: Harvard University Press.

Rawls, John. 1980/1999. “Kantian Constructivism and Moral Theory.” In *John Rawls: Collected Papers*, edited by Samuel Freeman: 303-358. Cambridge: Harvard University Press.

Rawls, John. 1993/2005. *Political Liberalism*. New York: Columbia University Press.

Rawls, John. 2000. *Lectures on the History of Moral Philosophy*. Cambridge: Harvard University Press.

Sensen, Oliver. 2015. *Kant on Moral Autonomy*. Cambridge: Cambridge University Press.

Taylor, Robert. 2011. *Reconstructing Rawls*. University Park: Pennsylvania State University Press.

Weithman, Paul. 2010. *Why Political Liberalism?* Oxford: Oxford University Press.

Weithman, Paul. 2017. “Autonomy and Disagreement about Justice in *Political Liberalism*.” *Ethics* 128: 95-122.