Philosophical Studies (2024) 181:23–41
https://doi.org/10.1007/s11098-023-02075-6

Stability and equilibrium in political liberalism

Paul Weithman[1]

Accepted: 27 October 2023 / Published online: 11 November 2023

Abstract
Threats to the stability of liberal democracies are of obvious contemporary import. Concern with stability runs through John Rawls's work. The stability that concerned him was that of fundamental terms of cooperation. Rawls long believed that the terms which would be stable were his two principles, but he eventually conceded that even a well-ordered society was more likely to be characterized by "justice pluralism" than by consensus on his own conception of justice. Contemporary liberal democracies, too, are divided about what justice demands. I believe Rawls's treatment of stability can help us understand the conditions under which fundamental terms of cooperation can be stable under non-ideal conditions such as ours. But because Rawls never worked through the consequences of his concession, his view needs to be developed before we can draw on it.

Rawls's treatment makes use of elementary game theory. Thus in *Theory of Justice* he said -- and in *Political Liberalism* he implied -- that stability would result from citizens with a sense of justice "playing" strategies which combined for what was, in effect, a Nash equilibrium. I argue that his concession requires a new conception of stability and implies that Rawls cannot appeal to a Nash equilibrium to show how stability would be maintained. His concession therefore forces open a troubling gap in his analysis. I fill that gap by proposing a weaker equilibrium concept that serves Rawlsian purposes. I conclude with what this project suggests for challenges facing the fragile liberal democracies of our own time.

Keywords Stability · Equilibrium · Political Liberalism · Rawls · Justice pluralism

✉ Paul Weithman
pweithma@nd.edu

[1] Department of Philosophy, University of Notre Dame, Notre Dame, IN 46556, USA

Threats to the stability of contemporary liberal democracies are of obvious import and concern.[1] One kind of threat is constitutional. Threats of this kind can arise when executives try to destabilize the balance of power set up by their societies' constitutions, and establish new and less democratic equilibria in which the balance of power between their own governmental branch and the legislature's is recalibrated so as to tilt in their favor.[2] These usurpations and encroachments would be worrisome whether or not they had popular support. A second kind of threat can arise -- even when a society's governmental forms remain basically the same -- if a significant number of citizens are willing to vote for candidates who will violate democratic norms, curtail voting rights through ostensibly legal means, leave the worst-off to make it as best they can or appoint judges willing to roll back fundamental rights. (See Mickey et al., 2017) These developments might not destabilize institutions, but they would destabilize a society's adherence to the principles it must satisfy if it is to be even a moderately just and effective liberal democracy.

A concern with stability runs through John Rawls's work from some his earliest essays (Rawls, 1999b, 96–116) to his last unpublished writings on political liberalism.[3] What concerned him was not institutional stability, but the stability of fundamental terms of social cooperation and its maintenance by a citizenry that acts on its sense of justice. Rawls's treatment of stability might therefore be thought to provide us with resources for understanding and addressing the second -- and arguably more pressing -- of the two threats I identified above. But Rawls's treatment is part of his ideal theory. The question he takes up is how a fully just society can remain so. *That* question, and Rawls's answer, may not seem relevant to the threats we face in our own far from ideal circumstances, when the democracies whose stability is under threat are far from fully just in the first place.

Such skepticism about the usefulness of Rawlsian ideal theory, widespread enough in political philosophy,[4] may be compounded by skepticism stemming from a concession Rawls made late in his development of political liberalism. In Rawls (1999a) and in much of Rawls (1996) -- hereafter '*TJ*' and '*PL*'[5] -- Rawls equated stability with the basic structure's enduring satisfaction of his own conception of justice, justice as fairness. But he eventually conceded that even a well-ordered society is likely to be characterized by pluralism about justice. (Rawls, 1996, 164) While that concession might seem to make Rawls's theory more realistic and less utopian, one prominent critic has said that, far from mitigating doubts about ideal theory, the concession throws the Rawlsian project into "disarray". (Gaus, 2016, 153) Others have reacted to the concession by saying that we should valorize a very different kind of stability

[1] I am grateful to Zach Barnett, Samuel Freeman, Alex Schaefer and Benjamin Straumann, and to audiences at the Indiana Philosophical Association, the Catholic University of Portugal and the conference on "Justice in the City" at the London School of Economics, for their helpful comments on earlier drafts.

[2] One scholar has called this kind of threat "executive aggrandizement"; see Bermeo, (2016, 10–13).

[3] For discussion of these writings, see Freeman (2023).

[4] Skepticism about ideal theory is registered in many quarters. Classic sources are Mills (2005) and Sen (2009). For a more recent example, see Wiens (2023).

[5] More precisely "hereafter referred to as, but not cited as, '*TJ*' and '*PL*'", since I will continue to cite the works by years of publication.

than Rawls studied (Thrasher and Vallier 2018) or -- in the extreme -- accept the fact that dynamism rather than stability is characteristic of a free society. (Schaefer, 2023)

On the contrary, I think that Rawls's analysis of stability can help us understand both the conditions under which fundamental terms of democratic cooperation can be stable even under non-ideal conditions, and the sort of stability which we who live in such conditions should take as our ideal. But I shall not argue for these claims here. Because Rawls never worked through the consequences of his concession, his view needs to be developed before it can be drawn upon or defended as the "realistic utopia" he suggested his well-ordered society would be. (Rawls, 2007, 11) In this paper, I show why its development is needed and I begin the task of developing it.

As we shall see, Rawls's treatment drew on elementary work in the theory of games. In *TJ*, we are told that stability would result from citizens with a sense of justice "playing" strategies which combined for what was, in effect, a Nash equilibrium. Rawls's transition to political liberalism was famously motivated by his recognition that *TJ*'s account of stability was inconsistent with the reasonable pluralism about the good that a well-ordered liberal society would encourage. For some time after that transition, Rawls continued to think that stability would be sustained by a Nash equilibrium. But as Rawls observed, recasting his view as a political liberalism required the introduction of a number of new concepts, including that of an overlapping consensus. And though Rawls did not note this explicitly, what he had to say about how the requisite equilibrium would come about was conjectural in a way that his earlier arguments had not been.

What I have called "Rawls's concession" -- his concession that a well-ordered society would be characterized by reasonable pluralism about justice as well as reasonable pluralism about the good -- meant that he had to weaken his earlier claim about what an overlapping consensus would be an overlap *on*. It would, he said, be a consensus on a family of liberal conceptions of justice rather than on justice as fairness alone. I shall contend that Rawls's concession, like his transition to political liberalism, requires significant conceptual innovation. For, I shall argue, it requires a new conception of stability. It also implies that Rawls cannot appeal to a Nash equilibrium to show how stability would be sustained.

Rawls's concession therefore forces open a troubling gap in his treatment of stability. I try to fill that gap by proposing a different and weaker equilibrium concept that would serve Rawlsian purposes. While much of this essay will be devoted to sustaining the game-theoretic reading of Rawls, getting Rawls right is not my ultimate aim. Rather, my aim is to introduce the concepts needed for an account of stability that is recognizably Rawlsian and that shares the attractions of the Rawls's own treatment, but that also accommodates justice pluralism. I conclude with brief remarks about what this project suggests for the challenges facing the fragile liberal democracies of our own time and place.

1 *TJ* and Nash

In Parts I and II of *TJ*, Rawls identified his two principles of justice using the original position and discussed institutions that would realize them. Part III of *TJ* takes up the question of stability. Rawls predicates 'stable' and its cognates of different subjects at different points in the book[6] but, as I have said, his fundamental concern is with whether a just society would remain just over time. That concern can best be appreciated by seeing the threat to stability with which Rawls was concerned. That threat can be stated in simple game-theoretic terms:

The fact that Rawls's principles would be chosen in the original position shows that everyone would be better off if they were complied with than not, relative to the relevant benchmark. It therefore shows, Rawls says, that the adoption of the principles is collectively rational. (Rawls, 1999a, 497) But we might still wonder whether conforming with the collective agreement on the principles is individually rational. If it is, then no one has sufficient reason to defect. An agreement from which no one has sufficient reason to defect is in equilibrium, so if conforming with the agreement reached in the original position is individually rational, then that agreement is an equilibrium point. If not, so that some or all would find it in their interest to defect from the agreement, then the agreement will come undone. The question of whether collective and individual rationality coincide is the question raised by prisoner's dilemmas. The threat to stability with which Rawls was concerned was, he avers, the threat of a generalized prisoner's dilemma. (Rawls, 1999a, 505) And so the stability question raised by Rawls can be put this way: would the agreement reached in the original position -- and hence the justice of society's basic structure -- be undone or undermined by such a dilemma? Or, instead, would the basic structure remain in a stable equilibrium on justice as fairness?

One way to stabilize an agreement is to attach sufficient penalties to defection, but Rawls prefers another solution. His solution falls into two parts.

- In Chap. 8 of *TJ*, Rawls argued that citizens who grow up in a society well-ordered by justice as fairness would acquire a sense of justice - a complex family of dispositions that includes a desire to act from Rawls's two principles of justice if others will also, and a readiness to make amends for acting unjustly.
- Rawls then argued, in Chap. 9, Sect. 86 of *TJ*, that citizens of a just society would each affirm a plan of life in which they treat their sense of justice as regulative.

To treat one's sense of justice as regulative is to give it categorical priority over other, competing, considerations. When faced with the opportunity to benefit from injustice, or to advance causes one cares about by acting unjustly, someone whose sense of justice is effectively regulative does not weigh those gains against her potential losses. For such a person, the possibility of acting unjustly is ruled out categorically, at least when she thinks others rule it out as well. And so in Chap. 9, Rawls argued

[6] For example, at 1999a, 154, Rawls refers to more and less stable conceptions of justice, at p. 350 he speaks of a well-ordered society as stable, and at p. 398 he predicates 'stable' of "a well-ordered society's conception of justice."

that each member of a just society would affirm a plan that gives her sense of justice this kind of priority, rather than trading off justice against other considerations case-by-case, and would take steps to preserve it.

Rawls says the arguments for this conclusion are meant to show that each would see that affirming his sense of justice as regulative "is his best reply to the similar plans of his associates". (Rawls, 1999a, 497). The phrase "best reply" is interesting and revelatory since it obviously calls to mind a Nash equilibrium. For a Nash equilibrium is a strategy combination in which each player's strategy is her best reply to the strategies played by all the others -- a strategy combination which is such that no one can do better for herself by playing something different, given what the rest are doing. The phrase "similar plans of his associates" is also important, for it confirms that Rawls thinks each will affirm his sense of justice as regulative provided others do. Thus what Rawls argues in *TJ*, Sect. 86 is that when others "play" plans of life that affirm their sense of justice, each person is best off affirming such a plan herself. The strategy combination in which each plays a plan of that kind is a Nash equilibrium. So let "the Nash claim" be Rawls's claim that affirming a regulative sense of justice informed by justice as fairness is each citizen's best reply to all others' doing the same.

A Nash equilibrium is, of course, a strategy combination which maximizes each player's utility function, given the strategies played by others. Rawls's arguments for the Nash claim do not refer to players' utility functions. Instead, Rawls makes assumptions about what citizens of a well-ordered society would most want to do, given what they know about what others most want to do. They would, he says, want to participate in a social union of social unions and to express their nature as free and equal rational persons. (Rawls, 1999a, 501–503) Since these are desires they can satisfy only if they have a regulative sense of justice, they would affirm that sentiment as regulative of their plans of life.

The Nash equilibrium of citizens' plans should not be confused with the equilibrium *institutions* are in when they satisfy a conception of justice over time. A society which satisfies a conception of justice is in a stable equilibrium just in case its basic structure enduringly or lastingly satisfies that conception. When the conception is justice as fairness, the Nash equilibrium of citizens' plans for which Rawls argues in *TJ*, Sect. 86 helps to explain the endurance. So the Nash equilibrium of citizens' plans is not identical to the institutional equilibrium on justice as fairness. Rather the Nash equilibrium helps to bring about, and so is distinct from, the stability of the institutional equilibrium.

Rawls does not say exactly *how* the Nash equilibrium of citizens' plans "translates" into stably just institutions. Presumably those who hold office in a well-ordered society, and those in positions to influence policy, support institutional compliance with justice as fairness. And presumably ordinary citizens do a well. And so when ordinary citizens must decide which candidates or policies to defend in public, and which candidates and policies to vote for, their choices will reflect their commitment to justice. Because each member's affirmation of her sense of justice is her best reply to others' affirmation of their sense of justice, their society remains well-ordered and just over time. The institutional equilibrium on justice as fairness will persist because of citizens' supportive attitudes.

That it persists *because of citizens' supportive attitudes* is important. I have argued elsewhere that Rawls's Nash equilibrium is self-enforcing. (Weithman, 2015) No outside enforcer, such as a Hobbesian sovereign, is needed to keep citizens from defecting from the agreement reached in the original position. That means that the justice of a well-ordered society is maintained by the free actions of its citizens, in a morally significant sense of 'free'.[7]

We have just seen that though Rawls's turns of phrase suggest that he thinks justice as fairness is stabilized by a Nash equilibrium, he does not use any of the technical apparatus his game-theoretic terminology leads us to expect. But when he introduces the ideas of stability and equilibrium, he observes that they are capable of mathematical refinement. (Rawls, 1999a, 400) It is therefore possible to state Rawls's arguments and conclusions somewhat more technically than he himself does:

According to Rawls's principles of justice, the justice of the basic structure depends on the extent of the basic liberties of the representative person, on equality of fair opportunity and on compliance with the difference principle. This suggests that we can represent the justice of a society in three-dimensional space. Let the x-axis measure the index of the representative person's basic liberties. There is a vast literature on the measurement of inequality of opportunity. (See Roemer & Trannoy, 2016) For now, I shall assume that there is a defensible measure and that it can be plotted on the y-axis. Some will want to plot economic inequality on the z-axis, perhaps using the Gini coefficient. But measures of inequality do not reveal whether what inequalities there are satisfy the difference principle, since they do not show whether those inequalities are to the maximum benefit of the least advantaged. Whether they are or not is, Rawls granted, bound to be controverted. And so the later Rawls granted that the difference principle could not serve as a constitutional essential. (Rawls 19,996, 229) A just constitution, he said, must guarantee a minimum. So let the z-axis measure the size of the guaranteed minimum.

There are presumably upper limits on what any society can do along the first two dimensions. While what societies can do along the third will vary depending upon their levels of economic development, all societies will face some limit or other along the third dimension as well. So the set of possible states of any society i will be a proper subset of the northeast quadrant of R^3. Call that proper subset J_i. Each of the points in J_i represents what we might call a society i's possible three-dimensional "justice states". One of those states, call it $<x_R,y_R,z_R>$, is the state that society is in when its basic structure satisfies justice as fairness. Following Alex Schaefer, let us suppose that there is a function f which tracks society's movement through J -- from one justice state to another -- over time.[8] Then Rawls's claim that justice as fairness would be stable is tantamount to the claim that a society which achieves justice state $<x_R,y_R,z_R>$ will remain there or will return to it in a relatively short interval after being perturbed. This claim, in turn, is roughly tantamount to the claim that

[7] In *TJ*, the relevant conception of freedom is autonomy; in *PL* it is political autonomy. I shall ignore those complications here.

[8] See Schaefer (2023), a paper to which I am deeply indebted; for my response to it, which shows the size of my debt, see my (2023).

$<x_R,y_R,z_R>$ is a fixed point of the function f. The Nash equilibrium of citizens' plans is what keeps a well-ordered society at that point.

2 Nash and the Early *PL*

In the 1980's, Rawls began to recast his view as a political liberalism because, he later said, *TJ*'s treatment of a well-ordered society's stability was "not consistent with [his] view as a whole." (Rawls, 1996, xviii)[9]. He continued to think of a well-ordered society as one well-ordered by his two principles, and he did not revisit the part of the stability argument in which he showed that citizens who come of age in such a society would have a sense of justice informed by the principles -- believing that the arguments of the relevant sections of *TJ* were basically correct. (See Rawls, 1996, 143 note 9) When he said that *TJ*'s treatment of stability was inconsistent, at least part of what he had in mind -- simply put -- was this. The assumption so crucial to his argument for the Nash claim -- the assumption that citizens of a well-ordered society would all value a social union of social unions and the expression of their nature -- was inconsistent with the moral pluralism encouraged by the liberal institutions of a just society.[10] To remove the inconsistency, Rawls needed a new account of why citizens of a well-ordered society would affirm their sense of justice as regulative, provided others would.

The new account, like the account of *TJ*, was a "best response" argument. In *PL* and related writings, Rawls identified certain political goods that can be had only by affirming a sense of justice informed by justice as fairness. One, for example, was expressing one's nature as a free and equal *citizen*. Rawls hoped that the reasonable comprehensive doctrines present in a well-ordered society would attach value to these goods. Because of the diversity of comprehensive doctrines, each doctrine might have to account for the value of these political goods for their own reasons. But if they did so, and if each attached sufficient weight to those goods, then Rawls could still argue that each member of the well-ordered society would judge that affirming a regulative sense of justice informed by justice as fairness was preferable to trading it off against other considerations case-by-case or rejecting it altogether. If those are the only strategies available to each citizen, Rawls could still argue that affirming a regulative sense of justice is each citizen's best reply to similar plans of others. He could then infer that office holders and citizens would act and vote as justice demands, and conclude that a well-ordered society would be stably just.

The central idea of this account of stability is, of course, the idea of an overlapping consensus of reasonable comprehensive doctrines. What those doctrines overlap on, in Rawls's initial presentations of his new account, are the political principles and

[9] For a superb treatment of Rawls's transition to political liberalism which draws on unpublished writings he produced during the period of transition, see Scheffler (2023).

[10] Thus in a little-noticed footnote in "Reply to Habermas", Rawls said that the idea of social union of social unions is "no longer viable as a political ideal once we recognize the fact of reasonable pluralism." (Rawls, 1996, 388 note 21).

values of justice as fairness. Because the focus of the overlapping consensus is a single conception of justice, the consensus is what we might call "single-focused."

Showing that such a consensus would obtain would vindicate the Nash claim since when such a consensus obtains, each adherent of each doctrine values "playing" justice as fairness. But recognizing wide-ranging, albeit reasonable, doctrinal pluralism seems to mean that there is little general information to be had -- few if any generalizations to be made -- about what citizens value. More precisely, there is little general information to be had about the inputs of citizens' utility functions. How, then, to argue that such a consensus would obtain? How, that is, to argue for the Nash claim, once we see the implications of reasonable pluralism?

Two expedients seem central to Rawls's answer:

- The first is to assume what Rawls calls "slippage" between comprehensive views and liberal principles, because most people's comprehensive views are only partial. (Rawls, 1996, 160) This leads to the optimistic conjecture that "many if not most citizens come to affirm the principles of justice incorporated into their constitution and political practice without seeing any particular connection, one way or the other, between those principles and their other views." (Rawls 19,996, 160).

What of citizens who *do* see a connection between their comprehensive doctrines and their political views? This brings us to Rawls's second expedient.

- Rawls conjectures that "a reasonable and effective political conception may bend comprehensive doctrines toward itself." (Rawls, 1996, 246) As a result, the reasonable comprehensive doctrines found in the well-ordered society of justice as fairness will come to endorse that conception of justice. His expedient is to "group [citizens] according to the doctrines they hold." (Rawls, 1996, 389) This allows him to conclude that citizens who adhere to those doctrines will endorse justice as fairness as well.

I am especially interested in the second conjecture: the development and liberalization of comprehensive doctrines. Unfortunately, Rawls provided only the barest arguments to support the conjecture. But it is worth observing what kind of argument could support it. The conjecture could be supported by looking for general features of reasonable comprehensive doctrines to see what makes them amenable to shaping or to liberalization. If there are no generalizations to be had about citizens' utility functions, perhaps there are generalizations to be had about their views of the good that could be used to support the Nash claim.

If the Nash claim is right, then everyone in a well-ordered society -- including those holding political office and those in positions to influence policy -- would support institutional compliance with justice as fairness. In that case, a single-focus consensus brings it about that their society will remain at $<x_R,y_R,z_R>$. And so as in *TJ*, so in much of *PL*, the Nash equilibrium of citizens' plans is supposed to be what explains or brings it about that $<x_R,y_R,z_R>$ is a stable equilibrium-state, the fixed point which a well-ordered society occupies in justice space. Thus as in *TJ*, so in

much of *PL*, Rawls's account of stability appealed to two equilibria, one of which causes the other.

3 The concession

But as Rawls continued to think through his view, he made what I implied at the outset is a significant concession. He conceded that the object of an overlapping consensus is less likely to be justice as fairness -- or any single conception of justice -- than a family of liberal political conceptions. (Rawls, 1996, 164) That is, he conceded that an overlapping consensus is more likely to be *multi-focal* than single-focused. This concession required Rawls to redefine some of the concepts that were central to his theory, and to introduce new ones.

One new concept is that of a liberal political conception of justice. Such conceptions have three defining features: They specify "certain rights, liberties and opportunities (of a kind familiar from democratic regimes)", they accord "a special priority to these freedoms" and they guarantee "all citizens, whatever their social position, adequate all-purpose means to make intelligent and effective use of their liberties and opportunities". (Rawls, 1996, xlviii) There are many liberal political conceptions which vary, among other ways, in the strength of the priority they accord to the rights and liberties, and in the threshold of adequacy they identify for the provision of all-purpose means. Some conceptions may permit the basic liberties to be more readily traded off for gains to the social minimum than others do. Some conceptions will set a higher social minimum than others. Rawls had previously taken a well-ordered society to be one whose basic structure satisfied justice as fairness. After the concession, he began to speak of a "well-ordered liberal society", by which he meant a society whose basic structure might satisfy different members of the liberal family, or hybrids of members of the liberal family, at different times.[11] (Freeman, 2023)

I take the fact that the family would be the focus of an overlapping consensus to imply that though the institutions of a well-ordered society may conform to different conceptions at different times, they will never conform with a conception of justice that is outside the family. "Keeping it in the family" is, we might say, the conception of stability required by Rawls's concession.

That conception can be made more precise with the spatial language used earlier. Different conceptions of justice for society i occupy different points in the three-dimensional justice-space J_i. I take Rawls's concession to imply that a society with an overlapping consensus may move from one point to another within the portion of the space containing liberal political conceptions, so that the function f that tracks i's movement through J_i need not have a fixed point. But though f need not have a fixed point -- though society can be well-ordered by different conceptions at different times -- Rawls conjectures that the class of liberal conceptions within the focus of an overlapping consensus would "vary within a certain more or less narrow range" (Rawls, 1996, 164) and that justice as fairness might be "the center of the focal class" (Rawls, 1996, 168). As I have observed elsewhere (Weithman, 2023, 205), we can take these

[11] I imagine that well-ordering by a hybrid is an outcome of political compromise.

remarks to mean or imply that members of the family are elements of a ball centered at justice as fairness with radius r[12] and that basic structures which are stably just will remain r-close to justice as fairness, for small r.[13]

4 Why Nash?

This new conception of stability complicates the problem of showing stability in ways that Rawls did not discuss, let alone resolve, for it fundamentally alters both the way the problem should be approached and the solution concept that should be used in approaching it. To see this, it helps to recall why Rawls argued for stability via Nash claims in *TJ* and early in the development of political liberalism.

In *TJ* and early in the development of political liberalism, Rawls wanted to show that:

(1) The basic structure of a well-ordered society would enduringly satisfy justice as fairness -- it would be in a stable equilibrium on justice as fairness -- and would do so for what he would later describe as "the right reasons".

To show that, it was necessary and -- given assumptions about how citizens' conduct affects the basic structure -- sufficient to show that:

(2) Ordinary citizens and those holding political office would freely act on justice as fairness, provided they think others will also.

To show that they would, it sufficed to show that:

(3) Each citizen would have and affirm a regulative sense of justice informed by justice as fairness.

And to show *that*, it sufficed to show that:

(4) Each would develop such a sense of justice in the normal course of moral development,

together with the Nash claim:

[12] Given a metric space (X; d) and a point p∈X, the open ball of radius r around p is $B_r(p) = \{q \in X : d(p,q) < r\}$.

[13] There is a further complication that I shall ignore here for simplicity's sake. In his latest writings, Rawls suggested that political liberalism should evince agnosticism about which liberal political conception is most reasonable; see Freeman (2023, 260). He thereby suggested, in effect, that political liberalism should take no position on which conception of justice is at the center of the ball. He thus seemed to suggest that political liberalism should treat reasonability, as it applies to conceptions of justice, to be what he earlier called a "range property". Interiority, as it applies to points inside a unit circle, is such a property because no point within a unit circle is any more interior to the circle than any other. Similarly, perhaps, no point in justice space that is within the ball of liberal political conceptions is any more reasonable as any other. For the notion of a range property, and its application to the unit circle, see Rawls (1999a, 444).

(5) Affirming as regulative a sense of justice informed by justice as fairness is each person's best reply to others' doing the same.

Thus the reason Rawls set out to establish the Nash claim (5) is that (1) depends upon (2). (2) poses a problem of strategic interaction among citizens and (5), together with (4), suffices to solve that problem. The support (5) helps to supply for (1), via (2), illustrates a point I made earlier: that one equilibrium supports another.

Once Rawls makes his concession, the problem of stability is no longer that of establishing (1). It is the problem of establishing:

(1') The basic structure of a well-ordered liberal society would enduringly satisfy one or another member of the liberal family, or a hybrid of members, and it would do so for what Rawls described as "the right reasons".

To establish (1'), it is necessary and -- again, given plausible assumptions -- sufficient to show:

(2') Ordinary citizens and those holding political office would freely act on one or another such conception of justice, provided they think others will also.

(2'), like (2), raises a problem of strategic interaction. But Rawls cannot argue for (1') via (2') in the way that he earlier argued for (1) via (2). For (2) was derived from (3), and the argument for (3) depended on (4) and (5). Once the possibility of reasonable pluralism about justice is recognized, however, it can no longer be assumed that all members of a well-ordered society have a sense of justice informed by justice as fairness. Instead, each citizen is taken to have a sense of justice informed by one or another conception of justice in the liberal political family. So what's true of a society with multi-focal consensus is not (4) but:

(4') Each citizen would develop a sense of justice informed by one or another member of the liberal family.

But if (4) cannot be established, then neither can (5), the other step from which (3) -- and hence (2) and (1) -- was derived. The possibility of getting to (1') from (2') shows that stability can still be established by solving a strategic interaction problem. The important question for present purposes is how that problem is to be solved.

The argument for (2) depended upon the assumption that citizens have just two strategies open to them: affirm one's sense of justice as regulative and decide case-by-case whether to act justly. Rawls took them to have just these two possible strategies because he thought the best way to support (2) was via (3). But while (3) is sufficient to establish (2), it is not necessary, and neither is the assumption of just two strategies. Another way to establish (2) would have been to suppose that citizens have as many strategies open to them as there are conceptions of justice on which they might act, and then to show that they would all act on justice as fairness. Whatever the merits or demerits of such an argument for (2), I believe the supposition holds much more promise for establishing (2') and (1'). Let me explain.

5 Why not Nash

In a society with a multi-focal consensus, citizens can all be supposed to have a sense of justice informed by one or another member of the liberal family. But members of a society with a multi-focal consensus live in a very different political environment than citizens who live in the well-ordered society of *TJ* or in a society with a single-focus consensus on justice as fairness. In societies with a multi-focal consensus, several -- perhaps many -- members of the liberal political family have their adherents and defenders. Even if conceptions of justice outside the family lack adherents, they may still be known and their eligibility for family membership may be actively debated.

The advocacy of all these conceptions of justice in the public and background cultures may make them attractive to those who do not hold them. And even if their advocacy does not make them attractive, their presence complicates citizens' strategic interaction. For example, faced with a majority's affirmation of a conception of justice other than her own, someone might think she could get more satisfactory political outcomes by advocating and voting for a compromise position than by insisting on what she believes justice demands. Moreover, political debate and political circumstances can change minds. In *TJ*, Rawls implies that people's conceptions of justice change slowly. (Rawls, 1999a, 498) If he is correct, then perhaps radical shifts in someone's sense of justice are impossible or unlikely. But if conceptions of justice in the multi-focal consensus do vary only in a "more or less narrow range", as Rawls said they do, then citizens can change from endorsing one conception to another without making changes that are radical. These changes, unlike those Rawls contemplated in *TJ*, might be made fairly quickly. Thus citizens might not affirm the same conception of justice at all times, and so might change the conception of justice on which they act even without compromising.

The possibilities of compromise and change mean that each citizen in a society with a multi-focal consensus has a large number of strategies open to her. Even if we assume that the number of conceptions of justice on which each might act is some relatively small number m, a population of n yields m^n possible strategy combinations. Identifying a Nash equilibrium or equilibria when citizens can act from any conception of justice in the political environment -- even any member of the liberal family -- would therefore require extensive knowledge of the distribution of political opinion and of other players' payoff functions, and the unwieldy comparison of a very large number of possibilities. Compromises and alterations may be hard to predict. The limited information available to each person, and the complexity of the computations involved, would make it very hard for each to anticipate the play of others so as to identify and play her own best response.

The difficulty of identifying and playing one's best response is compounded by a further consideration. I have assumed that if everyone affirms a conception of justice in the family of liberal political conceptions, then political outcomes will reflect one or another of those conceptions. But which of the conceptions they reflect depends upon processes of political decision-making and on the distribution of political power. Some people are better positioned than others to put their preferred conception of justice into effect because of the way decisions are made or because of the positions they hold. Determining one's best response to the strategies played by oth-

ers may requires her to anticipate the plays of those in different positions, and to anticipate officials' responses to strategies they expect the electorate to play. And so even if there is a unique strategy combination that constitutes a Nash equilibrium, the difficulties citizens will face in identifying their best response may make it unlikely that that combination will actually be played.

If that combination is unlikely to be played, then (2') and (1') cannot readily be defended by an argument that parallels one of the arguments for (2) and (1), and that depends upon the Nash claim. But the multiplicity of liberal conceptions -- and of strategies -- that complicate the attempt to defend (2') and (1') by appeal to a Nash claim open another possibility.

6 CURB and Congruence

To see the possibility, it will be useful recall how Rawls's original conception of stability could be made precise using the language of justice space, and how that conception was revised to apply to a multi-focal consensus.

Rawls's claim that justice as fairness would be stable is, I said earlier, tantamount to the claim that a society which achieves the "justice as fairness point" in justice space will remain there or will return to it in a relatively short interval after being perturbed. And this claim is tantamount to the claim that the justice as fairness point is a fixed point of the function f that tracks the well-ordered society through justice space. The well-ordered society's fixity at that point was brought about by the Nash equilibrium of citizens' plans -- hence by their plans' occupancy of a fixed point in their strategy space.

According to the revised conception of stability, a well-ordered liberal society is stable if it remains within a ball centered at justice as fairness. It need not remain at a single point. Once we see that society need not occupy a fixed point in justice space, there is no need to insist that citizens' strategy combination occupy a fixed point in strategy-space. And so there is no need for the combination to be a combination of best replies or to insist that the equilibrium concept that characterizes citizens' strategy-combination be point-valued. It can be set-valued instead.[14] For a variety of well-known reasons, a Nash equilibrium is often said to be too weak a solution concept. But in the current connection, it seems to be too strong. The question is: what kind of equilibrium among citizens' plans will stabilize a multi-focal consensus?

I now want to suggest that the appropriate equilibrium concept is one introduced by Kaushik Basu and Jörgen Weibull in a 1991 paper: the concept of a set of strategy profiles that is closed under rational behavior, or "CURB".[15] That concept is

[14] According to Voorneveld, Kets and Nordel (2005, 480 note 1), "a point-valued solution concept assigns to each game a collection of strategy profiles, i.e., a set of points in the strategy space of the game. A set-valued solution concept assigns to each game a collection of product sets of strategies, i.e., a set of product sets in the strategy space of the game."

[15] Basu and Weibull (1991). The concept might be called "closed by the common expectation of rational behavior", but the acronym "C-CERB" is less euphonious than "CURB". I am very grateful to Professor Basu for bringing his work to my attention and for suggesting that it might bear on a Rawlsian treatment of stability.

set- rather than point-valued, and is weaker or "coarser" than the concept of a Nash equilibrium. In a later work, Basu states the basic idea of a CURB set with admirable clarity.

When each player has a finite number of strategies open to her.

> a curb set is a collection of subsets of each player's feasible set of strategies, such that if each player believes that all others will remain within the specified subsets, she has no reason to want to employ a strategy outside her specified subset. (Basu, 2018, 64)

But Basu's basic idea, so stated, needs some elaboration.

First, according to this statement of the basic idea, the set that is CURB is a collection of subsets of each player's strategies. But it is clear from Basu and Weibull's technical exposition that CURB sets are *cross-products* or *sets of cross-products* of such subsets. For the sets that are CURB are sets of strategy profiles or strategy-combinations.

Second, I have supposed that the strategies facing citizens are conceptions of justice, each corresponding to a point in justice space. One of the coordinates of points in justice space is the social minimum. If, as economists sometimes assume, money can be treated as a continuous variable, then the social minimum can be varied continuously. If it can be, then the number of possible liberal political conceptions is, in principle, infinite rather than finite. I assume that citizens consider only a few conceptions and that the strategies are, practically speaking, finite. Even so, it will be helpful to elaborate the basic idea of CURB sets beyond the clear statement just quoted.

Let's start with citizens' strategies. There are lots of ways citizens might act on or "play" conceptions of justice, but for ease of exposition, I will restrict my attention to one of the actions I have referred to all along -- voting. For further ease of exposition, I make two simplifying assumptions. One is that citizens vote for conceptions of justice rather than for candidates who incorporate conceptions of justice into their campaign platforms. The other is that citizens act the way they vote, so that if they vote for a conception of justice, they do their part -- by way of obeying the law -- to uphold it.

Each citizen i's set of strategies S_i therefore consists of votes she might cast, one strategy for each conception of justice.[16] Strategy-combinations are n-tuples of votes, with n being the number of voters. Sets of strategy-combinations will then be sets of such n-tuples. Because each citizen i can vote for any conception of justice and not just members of the liberal family, votes for members of that family are subsets of all i's possible strategies. Call that subset X_i. Now consider the cross-product X of the X_i's. This is a set consisting of all and only the n-tuples in which all citizens vote for one or another member of the liberal family. My suggestion is that if that set X is a CURB set, then the multi-focal consensus on the liberal family will be stable.

One of the crucial steps in Basu and Weibull's definition of a CURB set is their characterization of players' beliefs -- in this case, beliefs about what ballots will be

[16] I shall suppose that all players have the same strategies open to them because any of them can vote for any conception of justice, though nothing turns on this assumption.

cast. Beliefs are taken to be probability assignments. I am interested in each player i's beliefs about, or the probabilities she assigns to, subsets of set X_{-i} -- subsets of that special set of (n-1)-tuples of votes in which all of the others vote for some liberal political conception. That is the set of (n-1)-tuples in which each person j other than i casts one or another member of her X_j. The case in which each citizen believes that every other person j will vote some subset of her X_j is the case in which everyone assigns probability 1 to everyone else's voting some subset of her X_j. So in that case, each citizen i assigns probability 1 to the set of (n-1)-tuples -- n-tuples less her entry -- being in a subset of set X_{-i}, and a zero probability to is being outside it.

Because I have supposed that citizens might change or compromise their views about justice, we do not need to suppose that any citizen i will respond in the same way to all combinations of votes. She might respond differently to different distributions of political opinion, thinking that she -- or the causes about which she cares -- will fare better if she tailors her views and votes to her circumstances. Assume there is a function that yields each voter's expected utility for any strategy open to her, given a belief about which subsets of X_j each of the others will actually play. Let $ß_i (X_{-i})$ be player i's optimal strategies when she believes that everyone else will vote for some liberal conception or other, bearing in mind that some of her own optimal strategies might not be in X_i -- bearing in mind, that is, that it's possible i's best response to liberal votes or sets of liberal votes cast by others *might* be a vote outside the liberal family.

Collect the best responses for all the players, yielding the set of sets of strategies $ß_1(X_{-1}) \ldots ß_n (X_{-n})$ for players 1…n. Now take the cross- or Cartesian product of the $ß_i (X_{-i})$'s. Taking the Cartesian product yields the set of strategy combinations or votes ß(X). This set consists of the cross-products of each person's best responses to the only sets of (n-1)-tuples of votes she thinks will actually be cast -- in the case of interest here, each person's best responses to sets of votes for liberal political conceptions. Of course, not every member of ß(X) is a sequence of votes each of which is a best response to all of the others in that sequence. One member might have player i voting for justice as fairness in response to a distribution of votes to which she would be better off responding differently. But it will include only sequences with this property: it will include only sequences of what each will do given what each expects others to do in response to what they expect her to do when we faced with a subset of X. The set X is CURB, or closed under rational behavior, if $ß(X) \subset X$.

If ß(X) is in X, then each person's best responses to anything he thinks others will actually do is in her X_i. In that case, no one will play anything outside her X_i and so no strategy combination outside X will be played. Another way to put the claim that the set X is CURB, following Basu and Weibull, is this: the *belief* that strategy combinations outside X will not be played implies that such combinations will not be played. For if, for each person i, i's belief that all others j will play a member of Xj leads her to respond with a member of Xi, then no strategy combination outside X will in fact be played.

I remarked earlier that the concept of a CURB set is a coarser solution concept than that of a Nash equilibrium. We can now see why. CURB sets need only contain best replies; they can contain much more besides. There are various ways of refining the concept by imposing conditions of tightness and minimality, but I shall not

discuss those conditions here. (Basu & Weibull, 1991, 143–44) What matters for present purposes is this:

If the set X is CURB, then no one has sufficient reason to vote for any conception of justice other than one or another member of the family of liberal political conceptions open to her, so the combination of votes that gets cast will always be a combination of votes for some liberal political conception. In that case, the overlapping consensus which has that family as its focus will be stable. For I have said the stability of a multi-focal consensus is not a matter of society's remaining at the fixed point in justice-space occupied by justice as fairness. It is a matter of its remaining within a ball centered at justice as fairness. Stability will be maintained if citizens play any combination of strategies that will result in society's remaining within the set or the ball. Given plausible assumptions about how citizens' votes are translated into political outcomes, society will remain within the set or the ball when citizens play any combination of liberal political conceptions.

Call the claim that they will "political liberalism's CURB claim". The suggestion for which I have been arguing, then, amounts to this. We have seen that once Rawls conceded the possibility of pluralism about justice, the stability claim he need to establish was not

(1) The basic structure of a well-ordered society would enduringly satisfy justice as fairness -- it would be in a stable equilibrium on justice as fairness -- and it would do so for what Rawls would later describe as "the right reasons".

but

(1') The basic structure of a well-ordered liberal society would enduringly satisfy one or another member of the liberal family, or a hybrid of members, and it would do so for what Rawls described as "the right reasons".

We have also seen that to establish (1'), it is necessary and -- given plausible assumptions -- sufficient to show:

(2') Ordinary citizens and those holding political office would freely act on one or another such conception of justice, provided they think others will also.

We have also seen that Rawls cannot establish (2') and (1') the way he earlier argued for (1), by appeal to a Nash claim. But, I suggest, he can establish (2') if he can establish political liberalism's CURB claim. For given the -- admittedly narrow -- way I have construed 'act' for purposes of exposition, that claim is tantamount to (2'). Since (2') is sufficient to establish (1'), the gap Rawls's concession opens in his stability argument is closed.

7 Conclusion

The stability secured by a CURB set has some nice properties.

Unlike the stability Rawls secured by a Nash equilibrium of citizens' strategy combinations, the stability secured by a CURB set does not require citizens to have any beliefs about how best to reply to any particular play by others. All they need to believe is that all others will vote for some liberal conception or other, and that when others do this, they are best off doing the same.

Like the stability Rawls secured by a Nash equilibrium, stability secured by a CURB set is self-enforcing. For what keeps each citizen from playing a strategy other than a liberal political conception are the common knowledge that no one has sufficient reason to play a strategy other than such a conception when others play a liberal conception, together with the common belief in rational behavior. That is why political liberalism's CURB claim is tantamount to the claim -- expressed by (2') -- that citizens act *freely* on one or another conception of justice and why stability secured by a CURB set is -- as (1') says -- stability for the right reasons.

Finally, like a Nash equilibrium, CURB sets can be shown to exist in games that satisfy certain conditions.[17]

But the fact that they can be points to a significant challenge for the Rawlsian project and for the attempt to draw conclusions from it for our world as it is. For when the Rawls of *TJ* argued that a just society would be stabilized by a Nash equilibrium, he did not show that the formal conditions for the existence of a Nash equilibrium were satisfied. He argued that citizens' desires were such that what each most wanted to do when others affirmed their sense of justice would lead her to affirm her own. Similarly, we might think, what really needs to be shown to establish the stability of a multi-focal consensus is not that the formal conditions that guarantee the existence of a CURB set are satisfied. What needs to be shown is that citizens' desires or utility functions would be such that they only want to vote for one or another member of the liberal family when they believe others will do the same. Unless we can do that, the concept of a CURB set serves only to label what still needs to be shown.

Unfortunately, I cannot address this challenge here. Instead, I shall comment on it very briefly.

We saw above that when Rawls tried to defend the Nash claim after his turn to political liberalism, he conjectured that reasonable comprehensive doctrines can be shaped or bent, and he employed the expedient of grouping citizens by comprehensive doctrine. Rawls's defense took that form because of the difficulty of making general claims about citizens' preferences in light of reasonable pluralism. It may be that defense of the CURB claim would be less reliant upon a conjecture of the sort on which Rawls previously relied. I do not deny the interest of studying reasonable comprehensive doctrines in order to identify features that allow for their liberalization and development. But I think we can learn something about why individual citizens would play strategies in their parts of the CURB set by looking again at the defining features of liberal political conceptions.

We have already seen that those conceptions have three features: they identify basic rights and liberties, they accord those rights and liberties priority of some kind, and they guarantee each citizen sufficient means to make use of her liberty. Concep-

[17] More precisely, CURB sets that satisfy the tightness condition can be shown to exist in those games; see Basu and Weibull (1991, 144).

tions with these features -- and only conceptions with these features -- satisfy the requirements of reciprocity, and do so in two ways. They are substantively reciprocal because they require social arrangements that are mutually beneficial. They are also what we might call "dialectically reciprocal" because each could reasonably offer them to others in the expectation that others could reasonably accept them. (Rawls, 1996, xliv) Thus, as Samuel Freeman has observed (2023, 261), for the late Rawls "reciprocity … becomes a basic idea". If citizens in a well-ordered liberal society all had a higher-order desire to live on terms that are reciprocal in these ways, (2') would be satisfied and that society would be stable in the sense of (1').

Rawls is sometimes accused -- as by Gerald Cohen (2008, 327 ff.) -- of being unable to acknowledge or address the fragility of justice. I believe that that accusation is mistaken. Even a nearly just society, one ordered by a liberal political conception, depends upon getting everyone in a pluralistic society to recognize the value of reciprocity. That they will is hardly a foregone conclusion. The injustices and polarization that mar our own societies only underscore that point. How citizens can be brought to desire reciprocity is an important question for civic education and public policy. Knowing what kind of stability a well-ordered liberal society would enjoy gives us an idea of the challenge we face.

References

Basu, K. (2018). *Republic of beliefs: A new approach to law and economics*. Princeton University Press.

Basu, K., & Weibull, J. W. (1991). Strategy subsets closed under rational behavior. *Economic Letters*, *36*, 141–146.

Bermeo, N. (2016). On democratic backsliding. *Journal of Democracy*, *27*, 5–19.

Cohen, G. (2008). *Rescuing justice and equality*. Harvard University Press.

Freeman, S. (2023). Reasonable political conceptions and the well-ordered liberal society. *Rawls's A Theory of Justice at 50* (pp. 257–276). ed. Weithman, Cambridge University Press.

Gaus, G. (2016). *Tyranny of the ideal*. Princeton University Press.

Mickey, R., Levitsky, S., & Way, L. A. (2017). Is the United States still safe for democracy? Why the United States is in danger of backsliding. *Foreign Affairs*, *96*, 20–29.

Mills, C. (2005). Ideal theory as ideology. *Hypatia*, *20*, 165–183.

Rawls, J. (1996). *Political liberalism*. Columbia University Press.

Rawls, J. (1999a). *A theory of justice*. Harvard University Press.

Rawls, J. (1999b). *Collected papers*. ed. Freeman. Harvard University Press.

Rawls, J. (2007). *Lectures in the history of political philosophy*. ed. Freeman. Harvard University Press.

Roemer, J. E., & Trannoy, A. (2016). Equality of opportunity: Theory and measurement. *Journal of Economic Literature*, *54*, 1288–1332.

Schaefer, A. (2023). Is justice a fixed point? *American Journal of Political Science*, *67*, 277–299.

Scheffler, S. (2023). Moral independence revisited: A note on the development of Rawls's thought from 1977–1980 and beyond. *Rawls's 'A Theory of Justice' at 50* (pp. 121–139). ed. Weithman, Cambridge University Press.

Sen, A. (2009). *The idea of justice*. Harvard University Press.

Thrasher, J., Vallier, & Kevin (2018). Political stability in the open society. *American Journal of Political Science*, *62*, 398–409.

Voorneveld, M., Kets, W., Nordel, & Henk (2005). An axiomatization of minimal curb sets. *International Journal of Game Theory*, *33*, 479–490.

Weithman, P. (2015). Relational equality, inherent stability and the reach of contractualism. *Social Philosophy & Policy*, *31*(2), 92–113.

Weithman, P. (2023). Fixed points and well-ordered societies. *Politics Philosophy & Economics*, *22*(2), 197–212.

Wiens, D. (2023). Against ideal guidance, again: A reply to Erman and Möller. *The Journal of Politics*, *85*, 784–788.