A Tale of Two Sets: Public Reason in Equilibrium*

Gerald Gaus

1 On Solving Two Problems of Public Reason Liberalism

Public reason liberalism is a family of theories according to which liberal political institutions, social structures and/or basic social rules are politically or morally justified if and only if they can be endorsed from the perspective of each and every free and equal "reasonable and rational" person.[1] Let us call these persons "the members of the justificatory public."[2] Public reason liberalism idealizes the members of the justificatory public in three senses. First, the members of the justificatory public are assumed to be free from at least some of the cognitive distortions and biases that often characterize actual people; we suppose that they generally reason in a sound way on the basis of relevant information. Different versions of public reason liberalism press this idealization quite far, while others insist on a "moderate idealization."[3] Secondly, the members of the justificatory public are idealized insofar as it is assumed that each is a good-willed person, concerned with living with others on terms that are mutually acceptable. In Rawls's language, we suppose that they are not simply rational, but "reasonable": they have a form of "moral sensibility" according to which they are ready to propose fair terms of cooperation, and are willing to abide by them "provided others can be relied on to do likewise."[4] The members of the justificatory public are thus idealized insofar as they are assumed to be moved by their sense of justice, or their aim to respect others as free and equal

* I have greatly benefitted from conversations with John Thrasher and Kevin Vallier. My thanks to them for these, as well as for allowing me read their insightful work-in-progress which deals with some of these matters.

persons. This leads to a third idealization: by assuming that the members of the justificatory public are moved by this moral sensibility, it is supposed that they do not pay attention to their reasons to ignore their sense of justice (or, we might say, defect on moral arrangements when doing so better advances their cherished ends, aims and projects). More generally, any specific version of public reason liberalism will hold that some reasons of actual people are not relevant to the justificatory question: the members of the justificatory public "bracket" (i.e., set aside) these reasons in their deliberations.

As Paul Weithman has shown in his recent study *Why Political Liberalism?*, these latter two idealizations raise a pair of problems for a public reason liberalism such as Rawls's.[5] First, we must consider whether once these last two idealizations are relaxed, actual well-reasoning citizens are apt to affirm that, all things considered, they have reasons to act on what was endorsed by members of the justificatory public. In the Rawlsian version of the problem we have to inquire whether a person will continue to affirm that she has reasons to act on the conclusions of the justificatory public once she takes up the "viewpoint of full deliberative rationality" in which she knows her full set of reasons, such as her conception of the good.[6] If not, a society regulated by the principles endorsed by the idealized justificatory public are likely to evince an instability: when people consider their full set of reasons they find the principles are not "fully" justified.[7] Let us call this *the problem of justificatory instability*. The second problem arises from the conditional nature of our moral sensibility. Assuming our first problem can be solved, we are prepared to act on just institutions only if "others can be relied on to do likewise." This, as Weithman points out, gives rise to an *assurance problem*.[8] Our sense of justice directs us to act on justified principles only if we can be assured that others will do so as well.

Weithman demonstrates that to solve these problems is to show how a society can reach a "just equilibrium."[9] The question I wish to explore here is *whether achieving just equilibria is facilitated or hampered by requiring that citizens share the same set of reasons, or the same way of reasoning*.[10] Rawls, we will see, originally thought that the problem of justificatory instability could be solved by showing that, from the perspective of "deliberative rationality," moral persons would endorse a common set of considerations that would lead them to affirm the dictates of their sense of justice. Weithman shows that Rawls's turn to political liberalism was based on the conviction that appealing to such shared reasons in our world of reasonable pluralism was unsustainable. Rawls thus developed a *convergence equilibrium model of justificatory stability*. I shall argue that not only was Rawls correct to do this, but once the convergence equilibrium model is in place, it largely supersedes the deep role of the argument based on shared reasons in the original position (though, of course, Rawls did not think so). I then turn to the second equilibrium problem, that of assurance. Weithman, and here he has recently been joined by Gillian K. Hatfield and Stephen Macedo,[11] hold that solving this problem requires, or at least is greatly facilitated by, citizens appealing to common public reasons. I shall argue that this is not so. Just as Rawls saw that the problem of justificatory instability in a world of reasonable pluralism can be solved by convergence reasoning, so too can the assurance problem be solved through each acting on those concerns that are relevant to her, but often not to others.

2 THE PROBLEM OF JUSTIFICATORY INSTABILITY

2.1 Justificatory Instability and the Gap Between the Idealized and the Actual

When deliberating whether to endorse some principle or rule *P*, the members of the justificatory public reason only on the basis of a subset of the reasons that an actual

person or citizen might draw upon. Picking up on a suggestion of Rawls, let us divide a person's evaluative considerations into two sets.[12] Call the set of *restricted* admissible evaluative considerations *R*, and the larger, essentially *unrestricted*, set of relevant reasons that are employed from the perspective of "deliberative rationality," *U*. Let us suppose for the moment that *U* contains *R*. The justificatory public, we assume, endorses principle *P* (and is conditionally willing to comply) on the basis of *R*, while actual people reason on the basis of *U*. The problem then arises that a person deliberating on the basis of *U* may no longer have a reason to endorse and/or comply with principle *P*. This is a greater problem the greater is the gap between the restricted set *R* and the unrestricted *U*. If *R* and *U* are not markedly different, then it would not be surprising if, especially on fundamental matters, a person is willing to endorse and act on the same principle given *R* and given *U*. We might say in this case that because the restricted set is not radically different from the unrestricted one, we expect people to generally confirm the conclusions of the justificatory public once the idealization is removed. However, when *R* and *U* are very different sets — if *R* is a *very* restricted set of reasons — then we may well encounter justificatory instability. The idealized justification does not, as it were, stand up in the full light of day, and people might not be expected to comply.

This problem clearly confronts Rawls's version of public reason liberalism as the set of considerations available to the members of his justificatory public behind the veil of ignorance is radically restricted, leaving a very large gap between *R* and *U*. Weithman draws our attention to a neglected discussion towards the end of *A Theory of Justice* in which Rawls considers the "hazards of the generalized prisoner's dilemma."[13] Since the principles of justice secure the basis of fair and efficient social cooperation, we prefer (even reasoning on *U* alone) that everyone acts on them rather than no one (remember, we always have reason to endorse them on the basis

of *R*). This much Hobbes taught us: "just institutions are collectively rational and to everyone's advantage from a suitably generalized perspective."[14] But supposing everyone else acts on them one may have reason (given *U*) to ignore one's sense of justice and free ride on the cooperative efforts of others.[15] Or, as Weithman points out, if actual individuals often conclude that acting on their sense of justice (being guided by *P*) consistently clashes with the advice of *U*, they "may resent their own sense of justice because of its costs. Even if they do not try to extirpate their sense of justice, they may well wonder what place it is rational for them to give that disposition in their plans of life."[16] Call this the *problem of ineffective endorsement*.

Even more worrisome is that, should the gap between *R* and *U* be large, once a person is aware of all her reasons in *U* — her conception of the good, her religious beliefs, her commitment to a comprehensive value theory, and so on — she may no longer endorse principle *P*. She may say, "yes, if I reason only on the basis of *R*, I endorse *P*, but once I tally up all the relevant reasons in *U*, the justification of *P* is 'overridden'."[17] Here the person is saying that the set of reasons *U*-minus-*R* (the elements of *U* not in *R*) contains a defeater of the justification for *P* from *R*.[18] In this case the person's reasons, all things considered, are to *not* endorse *P*. Call this the *problem of defeated endorsement*. So our first problem of justificatory instability takes two forms: the (sub) problems of ineffective and defeated endorsement.

These two forms of justificatory instability raise different issues. We may be tempted to say that the problem of ineffective endorsement is "merely" an empirical issue: it does not affect what is justified (in the sense of rationally endorsed) but "only" whether people will act on what is justified. This, though, draws far too sharp a contract between the justificatory and the empirical. One of the important lessons to be learned from Hobbes is that justified political principles plagued by widespread ineffective endorsement fail *as normative principles*, since they are unable

to help us solve the basic problems of social and political life. If a system of justice regularly confronts "generalized prisoner's dilemmas" that it cannot solve, it cannot serve as an effective basis of fair social cooperation for a society of free and equal persons. In this case, Rawls tells us, "the parties must reconsider the principles agreed to…."[19] We see here a complex interplay between the "justificatory" and the "empirical." Yet it is clear that this problem, while of justificatory relevance, is less deep than the problem of defeated endorsement. With defeated endorsement the justified principles are not ultimately to be rejected because they cannot perform a necessary function; rather, we see that in a broader context they simply are not justified at all.

2.2 The Congruence of R and U (The Double Shared Strategy)

To overcome these two (sub) problems of justificatory of instability, public reason liberalism needs to show that once a person reasons on the basis of the unrestricted set *U*, she will continue to affirm *P* (and so the problem of defeated endorsement is resolved) and she will have reason to act on *P* provided enough others do so as well (and so the problem of ineffective endorsement is overcome).[20] Weithman shows us that in *A Theory of Justice* Rawls focused on the problem of ineffective endorsement. The overriding aim in the third part of *Theory* was to show the feasibility of justice as fairness in the sense that the choice of principles in the original position can be "carried through."[21] The key to doing this, Rawls tells us, is to appreciate the congruence of the right and the good.[22] When we consider the good in terms of plans of life validated by deliberative rationality, we will see that our sense of justice is part of our good (and so we will not be alienated from it), and that because humans have "shared final ends" we see our participation in a just society as an expression of our nature.[23] This solution thus proposes *two shared sets of reasons*: the members of

the justificatory public share *R* which yields principle *P*, but stability is ensured by them also sharing, as it were, a core chunk of *U*, their unrestricted sets of reasons. This is what it meant by saying that shared ends are central to their notion of the good life. Reasoning on the basis of *R* leads to *P*, and reasoning on the basis of *U* affirms the justification of *P*, underwriting both our sense of justice and our tendency to act on *P*. "The hazards of the generalized prisoner's dilemma are removed by the match between the right and the good."[24]

2.3 Overlapping Consensus (on dropping one requirement of shared reasons)

As is well known, Rawls became convinced that this reply to the problems of justificatory instability was, for a variety of reasons, untenable.[25] The first solution contends that liberal institutions guided by *P* will be stable because members of the liberal society living under *P* will share core parts of *U*. However, Rawls concludes this ignores that "a plurality of reasonable yet comprehensive doctrine is the normal *result of human reason within the free institutions of a constitutional democratic regime*."[26] In short, life under a liberal *P* results in diversity of reasoning on the basis of *U*; even if members of a liberal society did share large chunks of *U* at some point, as they lived under liberal institutions this agreement would dissolve. Failing to share the unrestricted set of reasons *U* is endogenous to life under *P*, thus *P* cannot be stabilized by a core sharing of *U* (more specifically, the part of *U* that does not contain *R*).

Rawls thus comes to insist that "a democratic political society has no such shared values and ends apart from those falling under or connected with the political conception of justice itself."[27] This is to deny that significant sharing of the unrestricted set of reasons can be the source of stability. In his later work, then, Rawls replaces the consensus account of stability advanced in *Theory* with a

convergence account, according to which each person, on the basis of her own unrestricted set of reasons, affirms the justification of *P* on the basis of *R*. At least by the time Rawls writes the "Reply to Habermas" the problem of defeated endorsement has joined that of ineffective endorsement: Rawls is explicit that reasoning on the basis of *U* may override the justification of *P* on the basis of the "freestanding" argument from *R*, the restricted set of justificatory reasons. Indeed Rawls tells us that the argument from *R* is *P* is simply a "*pro tanto*" justification, which is only a "full" justification once a person confirms it on the basis of her unrestricted set.[28] Note that in *Theory* Rawls spends a third of the book explaining why the shared part of *U* endorses the "freestanding" argument; in his revised account this must be worked out by each citizen, since there are innumerable unrestricted sets that provide, hopefully, a variety of routes to endorsing the freestanding argument.

It is, I believe, a great mistake to maintain that Rawls in particular, or a revised account of public reason liberalism, can do without full justification based on *U*. Jonathan Quong, in his recent revision of Rawlsian public reason liberalism, explicitly denies that any further justification is required once the argument for the principles of justice (*P*) on the basis of the reasoning of *R* is completed. He writes:

> The objection [to the role of overlapping consensus, i.e. justification based on *U*] can be put in the form of a dilemma: (a) either the overlapping consensus is superfluous within political liberalism, since reasonable persons will *by definition* endorse the (correct) political conception of justice, or (b)… the overlapping consensus is not superfluous and people could (in the second justificatory stage) reject the political conception without being unreasonable. But if we embrace the second horn of the dilemma, this leads … to the … worry that people could veto the liberal conception of justice by claiming that

> it is not congruent with their illiberal views. If we want to preserve the liberal content of our theory, it is essential that such people are excluded from the constituency of the overlapping consensus. But can they be excluded in a way that does not also make the overlapping consensus superfluous to the justification of the political conception?[29]

We must be wary of justificatory victory by definition. Quong defines the reasonable in such a way that a person is reasonably only if she accepts that the argument from *R* to *P* is *conclusive*, and hence only unreasonable people would reject it. So by definition, any reconsideration of the justification in a wider context of reasons holds liberalism the hostage of unreasonable people. Now the case for *P* based on *R* is an inference on a very limited set of reasons. As liberals we need to know whether we have justified *P* by tailoring the set of justificatory reasons so as to ensure the justification of *P*, or whether this case stands up when the public — as real rational practical and epistemic agents must — consider their endorsement of the principles in the light of their various wider commitments. Suppose that a wide section of the citizens reject *P* when they consider *U*; to say that *P* is still upheld by all reasonable citizens is to ignore the fact that the liberal principle *P* simply fails to be justified to most citizens. The idealization (§1) is thus doing all the work; remove the idealization and the case collapses. What good can come of dismissing this as simply the objections of the unreasonable? It looks rather too much like dismissing as unreasonable anyone who fails to agree with us. At most (but see below, §2.5), we might say that a citizen is reasonable only if she endorses the case from *R* to *P*; if a person does not have reasons *R* (say, reasons based on conceiving of others as free and equal and being disposed to be fair to them), then we might say that she is not reasonable. ("Why should I enter the original position?" such a person might ask.

"I'm only concerned with number 1!"). So we might, without begging too many questions, say that a reasonable comprehensive doctrine is one that endorses the freestanding argument as a *pro tanto* justification. So it would still be an open question whether all reasonable comprehensive doctrines endorse *P* once citizens consider their wider sets of reasons. (This, I think, is something like Rawls had in mind.) Note that doing this would not, as Quong fears, hold justification "hostage" to illiberal views, if we mean by illiberal views those that reject the freestanding argument.

Leaving aside the stipulative nature of the argument, Quong's defense of liberalism is consistent with the liberal state requiring great coercion to maintain itself. Many citizens may simply have insufficient reasons to endorse the liberal state or act on the endorsement they would give on the basis of *R* alone. Because the liberal principle would not appeal to the reason of many it is likely that it could only be upheld by the oppressive use of coercion, "with all its official crimes and the inevitable brutality and cruelties."[30] To the extent liberalism fails to resolve the problem of ineffective endorsement, widespread state coercion will be required to stabilize liberal practices; to the extent that the problem of defeated endorsement cannot be solved, in the name of reasonableness the liberal state will force its citizens to act against their consciences. Such a regime will not be enduring or secure.[31] It constitutes a liberal authoritarianism.

2.4 A Free Justificatory Equilibrium

In a social world characterized by deep pluralism, Rawls came to realize that justificatory stability could only be achieved by a justificatory equilibrium in which citizens reasoning on their unrestricted set of reasons affirm, and tend to comply with, basic principles and institutions. Such an equilibrium can be contrasted to

justificatory stability via a shared social ideal. In a society stabilized by a shared social ideal, we affirm and act on our core moral and political rules, principles, and institutions for the same reasons: we share a common outlook or public way of life that stabilizes the social order. Such an order partakes of a community or an association: our shared social ideals and their shared basis unite us, and as participants in the social and political order what separates us is simply not relevant. We are, one might say, first and foremost citizens. To use Durkheim's famous term, stability is achieved by a "mechanical unity" based on commonality.[32] As social thinkers from Durkheim to Rawls realized, modern pluralism undermines stability based on such unity. In a deeply pluralistic world justificatory stability can only be achieved if endorsing and conditionally complying with the basic principles are each person's "best response" (determined by her *U*), to the endorsement and conditional compliance by others. We act the same way for different reasons. Justificatory stability is achieved by a sort of Nash equilibrium.[33] In such an equilibrium there is no need to bracket our differences and so base our social life on shared ends; we draw comprehensively on our reasons and determine whether our unrestricted set of reasons instruct us that just action is the best response to the just action of others. As I have argued elsewhere, such a Nash equilibrium is a genuine expression of our freedom as agents in a social world.[34] It is the freedom of a social agent, in a world of other agents, to act as he thinks best given the legitimate actions of others.

A great benefit of stability through such an equilibrium is that, in stark contrast to stability via shared reasons, it can cope with the indeterminacy of our reasoning based on *R*. As is well known, while in *A Theory of Justice* Rawls believed there was an unequivocal best liberal theory of justice ("justice as fairness"), he came to believe that there is a set of reasonable liberal views.[35] Suppose then that, employing their reasoning the best they can, members of the justificatory public

arrive at the conclusion that principles P_1 or P_2 are better than all alternatives, but they cannot agree on a ranking. Without having shared reasons for either alternative, the shared reasons view appears to provide no clue as to how we could arrive at a stable equilibrium on either.[36] Appealing to *U* is supposed to allow us to equilibrate on the freestanding argument from *R* to *P*, but here we have arguments from *R* to P_1 and from *R* to P_2. Inference on the basis of *R* cannot help.

What can help is the idea of a best response on the basis of *U* to other people's actions. We are confronted with an impure coordination game as in Display 1 (higher numbers indicate more preferred outcomes). Here the differing results of the freestanding argument are ordered by each on the basis of their unrestricted set *U*. In Display 1 Alf and Betty disagree on which is best supported, but concur that that the best response to the other adopting a principle is to also adopt it.

		Betty	
		P_1	P_2
Alf	P_1	2, 1	0, 0
	P_2	0, 0	1, 2

DISPLAY 1

A one-shot two-person game can give us some insight, but it is clearly an inadequate way to model the selection of a particular member of the set of principles justified by *R*. The relevant coordination problem is not a single-play game, but an iterated game. We have a number of encounters with others, and each can be understood as a play in a series of impure coordination games over many options. Now in an iterated game a person's utility is a combination of her utility in this play, plus her expectations for utility in future games. Thus a person might sacrifice utility in one play to induce play in future moves that will yield her a more favored result. Now in

large iterated games a bandwagon effect manifests itself. As I have argued elsewhere, such large-person iterated coordination games exhibit a strong increasing returns effect: the more people come to embrace a particular rule, the more reason others have to also embrace it.[37] In a wide range of circumstances a society can come to coordinate on a stable outcome even given the indeterminacy of the freestanding argument if people can draw on *U*. Note here that *U* plays a crucial justificatory role in completing the justification of the principles of justice.

2.5 On Dropping All Requirements of Shared Reasons

Of course to say that a free justificatory equilibrium is the only plausible device of justificatory stability in a deeply pluralistic world (in which we do not privilege a common thick set of shared values in our social and political relations) does not show that such an equilibrium can be achieved. If it cannot, the project of public reason liberalism fails: no stable political order among free and diverse people is possible (at least not one without a great use of coercion). It is thus a fundamental desideratum for public reason liberalism to facilitate the rise of a free justificatory equilibrium.

Consequently, the most plausible version of public reason liberalism must seek to maximize the prospects of a free justificatory equilibrium. Given this, we must inquire whether we should reconsider the requirement that all accept the freestanding argument based on the restricted set of reasons, *R*. The view we have thus far been considering advances two requirements for a justificatory stability:

(1) *P* is justified only if it is endorsed on the basis of the shared restricted set *R* ("freestanding justification").

(2) The freestanding justification of *P* based on *R* must be affirmed by citizens when appealing to their unrestricted set *U* ("full justification").[38]

Let us consider four groups of citizens in relation to these two requirements:

A. Those for whom both (1) and (2) apply.

B. Those for whom only (1) applies.

C. Those for whom only (2) applies.[39]

D. Those for whom neither applies.

Quong, we have seen, holds that *P* is adequately justified for groups A and B; we might say that Quong's strategy is to maximize the population that endorses the principles by restricting all justification to the restricted set, *R*. But, as we have seen, this invites justificatory instability. Rawls seeks to avoid justificatory instability by holding that only for group A is *P* fully justified. Note that both Rawls and Quong hold that *P* is not justified to group C, even though when group C considers their unrestricted set of reasons *U*, they affirm it. So even though this group sees principle *P* as justified given their all-things-considered judgments, they are excluded by both Rawls and revisionists such as Quong from population to whom it is justified. Given the importance of solving the problem of justificatory instability, excluding these citizens simply because they have not accepted the canonical *R*-based argument looks myopic. Such public reason liberals are making it harder to achieve a free stable order.

We are led to the suggestion that the necessity of the shared freestanding argument from *R* to *P* be dropped. To be sure, we may still employ it in our exposition of the case for liberalism; to the extent we share reasons, that is all well and good, and the *R* to *P* argument may give us insights into the liberal justificatory project. But we do not *require* that a person accepts the case from *R* to *P* in order for *P* to be justified to her. Note that by basing justification on *U* (or some subset, *U** that is rather close to *U*), we pretty much eliminate the problem of justificatory

instability, since the justificatory set of reasons approaches the all-things-considered set.[40]

The Rawlsian might advance two worries about this expansion of the set of reasons to be considered by the justificatory public. First, they are apt to worry that their favored principle *P* (say, the difference principle) will fail to be justified given this wider set. This is manifestly Quong's concern: he is deeply concerned that "illiberals" will veto *P*, which is why he insists that only the support of group B is necessary. Rawls — and here I have concurred — cannot take heart at this victory, for whatever justification achieved by only admitting B as the relevant public is apt to incur a high cost in justificatory stability. Once we see that justificatory stability requires (2), why also insist on (1)? Again, I suspect that the Rawlsians' worry is that we will get the "wrong" results: once we abandon the two-staged account, we may achieve a more comprehensive equilibrium on some alternative principle *P**. The two-stage account, like any elimination process, is path-dependent: an alternative *P** that may be more favored by citizen's unrestricted sets (*U*) can be eliminated at the first stage based on *R* alone.[41] Surely, though, we do not wish a moral and/or political order among free and equal persons to settle on a certain principle because we have devised a path-dependent justificatory process that eliminates non-liberal (or at least non-Rawlsian) competitors at an early stage, though they would be endorsed in a wider setting. As liberals, we wish to confirm our conviction that liberal principles can be freely endorsed by all free and equal persons (or citizens), not show that we can devise a path-dependent procedure whereby they are selected. If a wider body of citizens freely endorse *P** once they consider all that is relevant, what case remains for insisting that *P* is *really* the principle that is justified to all?

2.6 Interpretive Equilibria

To say that public reason liberalism should jettison the requirement that justificatory reasons be shared does not mean that nothing must be shared among actual citizens. We must share interpretations of the justified rules, principles and institutions. If the rules, principles and/or institutions are to structure a common cooperative social life, we clearly must entertain similar understandings of the rule, or look to the same procedures (such as the courts) to resolve any disagreements that arise. As Cristina Bicchieri stresses, for effective coordination via norms, rules, etc. participants must share "scripts": shared expectations about what is called for in various circumstances.[42] Norms, rules, and institutions can be understood as constituting "correlated equilibria" in which individuals focus on a common signal (the norm, rule, law) to "choreograph" their actions. Thus, for example, to "share" a property rule is to share a complex interplay of expectations about what a property owner will do when another trespasses on her land, or what signs one can post on one's buildings, and what will happen if one posts unacceptable signs.[43] We can explain the development of such correlated norms without supposing that the individuals share ends or reasons to endorse the norm. (Fred Astaire and Ginger Rogers need not have had common reasons for dancing together.) It is essential not to conflate this choreography inherent in all norms — our dance of first-order expectations — with shared public justificatory reasons.[44]

3 CONDITIONAL COMPLIANCE AND THE "ASSURANCE PROBLEM"

3.1 The Assurance Problem and Rawlsian Public Justification

Suppose we succeed in fully justifying principle (or rule, or institution) *P* to all free and equal persons on the basis of convergence reasoning alone (that is, we drop all shared reasons requirements for full justification). Suppose further — as I have been

arguing — that this best solves the problems of ineffective and defeated endorsements. Now Rawls seems quite right that the justification for *P* is still conditional: our citizens are willing to comply with *P* on the assumption that others do so as well. Thus, as Weithman points out, citizens face an assurance problem, a very simple version of which he models in a 2×2 game, as in Display 2 (again higher numbers indicate more preferred outcomes).

		Betty	
		Act on P	*Do not*
Alf	*Act on P*	3 3	2 0
	Do not	0 2	1 1

DISPLAY 2

In this assurance game, Alf and Betty both receive his/her highest payoff if they both act on *P*. However, Alf gets his second highest payoff if Betty acts on *P* while he does not (Betty acts fairly while he acts to advance his concerns in an unrestricted manner). Betty's reasoning is symmetric. Each ranks being the unilateral follower of *P* as the worst outcome. Thus universal defection (no one acting on *P*) is preferred by Alf to unilaterally acting on *P* (and Betty prefers it to her unilaterally acting on *P*). This game has two Nash equilibria: both acting on *P* and the Pareto-inferior neither acting on *P*. We call this the "assurance game" because, unless Alf and Betty can be assured that the other will play cooperatively, they will end up in the Pareto-inferior equilibrium.

> Weithman writes that this threat of instability
>
> arises on the assumption that each person wants to act justly, but needs the assurance that he will not be taken advantage of. Since a WOS [well-ordered

> society] is a just society, everyone is already behaving justly, so what each person needs to be assured of is that others will continue to act justly rather than defect. Suppose that each person knows everyone else's balance of reasons tilts in favor of acting justly when others do…. Then each knows that no one else has sufficient reason to take advantage of him and the *mutual assurance problem* is solved.[45]

Thus, Weithman argues that "public knowledge of an overlapping consensus is therefore sufficient to solve the *mutual assurance problem*."[46] This knowledge, Weithman and others hold, is conveyed through what Rawls identifies as the third stage of justification — public justification. After the *pro tanto* justification of the free-standing argument and the overlapping consensus of full justification, comes public justification, which

> happens when all reasonable members of political society carry out a justification of the shared political conception by embedding it in their several reasonable comprehensive views. … A crucial point here is that while the public justification of the political conception depends on reasonable comprehensive doctrines, it does so only in an indirect way. That is, the express contents of these doctrines have no normative role in public justification; citizens do not look into an account of others' doctrines, and so remain within the bounds of the political.[47]

Public justification so understood appears to be a public knowledge that overlapping consensus has been achieved. For Rawls and his followers such as Weithman and Macedo[48] this is achieved by a public political culture that restricts itself to some ideal of public reasoning as shared reasoning on the basis of the justified political conception.[49] By constraining their public reasoning to the shared political conception (let us call this "a display of shared public reasoning"), it is

supposed, citizens assure each other than they have embedded this political conception in their comprehensive doctrine, and so they assure each other that it is justified in the unrestricted set *U*. Note that, if this is so, it demonstrates why we cannot do without the two-staged argument I criticized in section 2.5. Once we have solved the problem of justificatory instability by embedding the argument from *R* to *P* in our unrestricted sets, we solve the assurance problem by, essentially, only appealing to the *R* to *P* argument (and the ways of reasoning it licenses)[50] in certain fundamental political discussions. Thus those who endorse *P* on the basis of *U* but not on the basis of *R* cannot help solve the assurance problem.

3.2 Why a Display of Shared Public Reasoning Will Not Solve the Assurance Game

There is an insight at the heart of the Rawlsian argument: overt communication within a group that indicates allegiance to group norms can indeed increase a tendency to trust each other. We can understand Rawlsian displays of shared public reasoning as what economists call "cheap talk" — and that can positively influence cooperative behavior.[51] We certainly should accept the insight that a display of shared public reasoning might be one way to convince others that one is a trustworthy citizen, but there is no reason to suppose that it is a unique, or indeed especially effective, way to do so. Displays of allegiance to the political system, affirmations of the importance of upholding the law, all may serve the function of assuring others of our propensity to cooperate. Whether conducting arguments — and so *dis*agreeing — in a certain constrained language is an effective way to display trustworthiness is controversial. In these contexts we are disagreeing with others, and so the display of shared public reasoning will also be a display of disagreement. We are sending a mixed message of agreement *and* disagreement; sending such messages may not be the most effective way to establish trust. If the main claim is a

psychological one about what types of communications and displays are apt to induce trust in very large groups, then the matter depends on the psychological evidence. For now, let us grant that some sorts of public displays are of use in engendering mutual trust. However, I will argue in section 3.3 that they are of secondary importance in our political context.

If we move from general psychological claims about inducing trust to the more technical problem of the assurance game, displays of shared public reasoning will not suffice to solve the problem, and this for two reasons. (*i*) In the assurance game in Display 2, Alf's communication that he intends to cooperate will not help Betty form beliefs about what Alf will do, because regardless of what Alf intends he would have a reason to send an "I will cooperate" message.[52] If Alf plans to cooperate he would tell Betty so (since he will get his best payoff if she also cooperates), but if he plans to defect he *still* has reason to give a display of trustworthiness, since he prefers his unilateral defection to mutual defection, and so has an incentive to induce Betty to cooperate. So in Display 2 communication of one's readiness to comply does not help the other form beliefs about your intentions.[53]

(*ii*) It is not enough to solve the assurance problem that "each person knows everyone else's balance of reason tilts in favor of acting justly when others do." We need something considerably stronger: we need *common knowledge* of this fact.[54] Suppose that I know that everyone else's balance of reason tilts in favor of acting justly when others do, but I am not sure that others know that I know this. That we each know X does not imply that we each know that we each know X. I not only have to be sending the signal that I endorse the "*R* to *P*" reasoning but I have to know that the others are properly receiving my message and, so, for example, do not infer from the fact that I am disagreeing within them on the basis of shared public

reasons that I am simply trying to manipulate them, and will defect if I do not win the debate. And, in turn, should they know that I am not a defector simply using shared public reasoning, they must know that I know that they know I am not a defector. Common knowledge is a very strong assumption; as Hatfield and Macedo recognize, it implies a common knowledge of each other's logicality as well as information. But we are seldom in a world of such knowledge; a solution to the problems of large-scale assurance and coordination that depends on it cannot be convincing.[55]

3.2 Thinking about Stability Under Conditional Compliance

Even though thinking through the issue in terms of a simple one-play assurance game is not helpful, the problem of conditional compliance is real. In a plausible account of norm following we have to suppose that (*i*) even if a person has overall reason to follow *P*, (*ii*) she only has reason to act on *P* if enough others follow *P*.[56] It is important to stress that, while there is a role for the normative expectations of others in explaining why we comply, a critical factor is our first-order empirical expectations about what others will do. There is sound evidence that a person's first-order empirical expectations about how others actually act ("do people comply with the norm?") is a powerful explanatory factor in explaining whether a person will comply. Indeed, the evidence indicates first-order empirical expectations are a much more powerful factor than normative talk ("This is our norm, which I affirm").[57] I stress that these are first-order expectations: they are one's beliefs about what others will do, not my beliefs about their beliefs about my beliefs..., which are required for common knowledge solutions to compliance problems.

To begin to see how we might explain the evolution of conditional cooperation let us consider some very simple dynamic models. Call β_i person *i's*

threshold level as to what is "enough" compliance: suppose that β varies between 0 (in which case the person is a unilateral complier) and *N*-1, in which case the person will only comply once everyone else is complying. There is no reason to think that all have the same β values. Some may have a β of 0, being essentially unconditional compliers.[58] Let us assume a roughly continuous range of β values, perhaps with some normal-like distribution around mid-range values. Important in modeling the evolution of compliance would be information about what others are doing. Let us consider two cases: evolution under perfect/near-prefect information about the actions of all others, and only local knowledge about what others are doing.

3.2.1 PERFECT INFORMATION. If we assume that each and every person has full knowledge of the compliance of others, and continuous range β values, we can see how iterated interactions can lead to full compliance. Starting with the unconditional compliers in the first round, those with β greater than but near 0 would then have their threshold met, and so on, ending with compliance by the person from whom β = *N*-1 person. As we leave the tail ends of the distribution and approach the middle the process would speed up, and again slow down as we near the further tail. Call this the *compliance cascade*.

While the compliance cascade is easy to envisage under these conditions, it is subject to two possible problems under less-than-perfect information: a reverse cascade and "stalled" cascades. Suppose that we have reached full compliance: as does Rawls, we assume a well-ordered society with full compliance. Suppose now that the person for whom β = *N*-1 mistakenly comes to believe that another has failed to comply; if so (*mutatis mutandis* by the reasoning above) that error will lead to a reverse cascade to zero compliance (assuming that there are no countervailing

errors along the way that blocks the reverse cascade).[59] However, this reverse cascade depends on a number of assumptions. It will be thwarted if the highest β values are well short of *N-1*. That is, if we suppose that the top β values are, say, .9, then even a number of mistaken judgments of non-compliance will not unravel the well-ordered society. Especially if we begin with the assumption of full compliance (rather than having to explain how it comes about), plausible distributions of β allow for the stability of first-order expectations and so fulfilling the conditions required for conditional compliance.

Allowing for mistaken judgments about the compliance of others can also *stall* the cascade: the person who, as it were, should now be ready to comply mistakenly thinks that her threshold has not been met, and so refuses to comply, perhaps halting the cascade. Interestingly, this would be a problem early and late in the process; when we are at the tail end of the distributions, the mistakes of a few people could either stop the process from getting going, or halt it short of full compliance (assuming again that there are some for whom approximately $\beta = N\text{-}1$). In the middle of the distribution, where many people have the same threshold, we would not expect small mistakes to have such consequences. And this seems correct: intuitively it is easy to see how the cascade may have a hard time getting going or completing itself.

3.2.2 LOCAL KNOWLEDGE. Usually we only know what those around us — those with whom we have opportunity to interact — are doing. Under these conditions the dynamics of assurance and compliance, not surprisingly, are much more complicated.[60] Let us start with the Rawlsian problem: will a society of full compliance be stable? Again we need to suppose that compliance is a first-order

expectation about what people will do: we have knowledge of the norm, and so can detect who is cheating (or, at least, not complying).[61] In this case breakdown of compliance may be caused either (1) by some citizens withdrawing their support of the rule such that for the mass of citizens actual compliance falls below their β value and/or (2) mistaken judgments that (1) is occurring, which cause such a breakdown. Suppose then that person *i* only can form judgments about what is going on in some group of neighbors *H*. A Rawlsian would say that stability can be achieved within *H* by people advocating the public conception of justice, and stressing their devotion to the *R* to *P* freestanding argument as well as the principles of public reasoning it establishes. It is not clear how helpful this will be. The worry about the stability of the equilibrium only arises once a citizen observes (1), or makes the mistake of (2), in her *H*. If observed violators continue to give the public message affirming the public conception, citizen Betty is confronted with, as it were, talk ("I, Alf, affirm *P* on the basis of *R* given my *U*") that clashes with observed behavior (Alf just violated *P*). Given Alf's incentive to affirm his endorsement of *P* regardless of whether he has, or intends to conform (§3.1), Betty cannot much rely on such talk for evidence about Alf's compliance. Surely it will be Alf's behavior that will be crucial in making her judgment.

On the convergence account there is no canonical argument for *P;* whether P is justified is a matter of whether a citizens' unrestricted sets endorse it. As a convergence reasoner, in observing her neighborhood (*H*) Betty will be concerned with observed rates of defection. Supposing that *P* is in equilibrium with citizens' unrestricted set of reasons, then rates of defection should be low. Mistakes about the rule, and mistakes about whether others have conformed to the rule, will lead to some baseline rate of perceived defection in *H*. So long as within the neighborhood

H this baseline rate of perceived defections does not drop Betty below her threshold of compliance (β) she will continue to conform to *P*. It is thus crucial for stability that few citizens have β values approaching 1; such values render the justificatory equilibrium susceptible to "trembling hands" — mistakes and errors about compliance.

However, there are bound to be outlier neighborhoods: those in which non-compliance is high (perhaps because of unusually high β values). A resident of this neighborhood will form pessimistic estimates of overall compliance. She (and others like her in *H*) may be driven below their threshold, and so also cease to comply with *P*. Suppose, then, that everyone in neighborhood *H* ceases compliance; note that *H* borders other neighborhoods (say $H_1 \ldots H_8$) as in Display 3.

H_1	H_2	H_3
H_4	H	H_5
H_6	H_7	H_8

DISPLAY 3

We suppose that those at the edge of a neighborhood interact with, and so know about, those on edges of adjacent neighborhoods. Now that *H* is a non-compliant neighborhood, the adjacent neighborhoods all will have significantly increased their interactions with non-compliers. This raises the troubling possibility that their interactions with *H* may push some or all of them below their compliance thresholds spreading out non-compliance.[62] Whether this occurs will, of course, depend on the

β values of the members of the other neighborhoods as well as the initial perceived rates of defection in them; but we do know that interaction with *H* will tend to push down perceived compliance in these adjacent neighborhoods.

Perhaps the most effective way to check the danger of such a noncompliance epidemic is through punishment.[63] We have been assuming that Betty is merely reactive: she establishes her estimate of compliance and checks to see whether it meets her threshold level. Should noncompliance increase, Betty may find that her neighborhood is now below her β value, and so she herself ceases compliance. However if Betty is a "Rule-following Punisher,"[64] she not only has a conditional tendency to follow rules (or norms, etc.), but she is willing to forgo some resources to punish those who do not comply, thus stopping an epidemic of non-compliance. This sort of decentralized enforcement helps to counteract non-compliance, thus stabilizing norms in the face of temptation to defect.[65] It is very hard to see how stability can be secured in the face of imperfect information without willingness of many to punish perceived violators.

4 CONCLUSION

The move from *A Theory of Justice* to *Political Liberalism* was characterized by a conviction that a stable equilibrium on justice could be achieved without citizens sharing a great part of their unrestricted set of reasons. What I have called the "double sharing" strategy of *Theory* (§2.2) was abandoned by Rawls in favor of (to of course simplify an exceedingly complex corpus) a single-shared view (§2.3), in which we share the freestanding argument, but fill out the full justification of the principles in different ways. I have argued that the best prospect for a stable equilibrium on justified rules and principles is to drop all requirements of shared

justificatory reasons (§2.5) Such an approach has the best prospects of solving the problems of ineffective and defeated endorsements, and, *pace* Rawls and some of his followers, there is no good reason to think that a public display of a shared conception of justice is needed to show (or even is particularly helpful in showing) why conditional compliers will become, and remain, actual compliers (§3). Rawls's followers may well be worried that if we drop the canonical shared reasons requirement (underlying the freestanding argument) we can no longer be guaranteed that liberal principles will be the core of a justified equilibrium. However, the fundamental commitment of those devoted to a free social order is that our social rules, norms and principles must be a justified and stable equilibrium. I have faith that those will be the fundamental liberal principles — but that, hopefully, is the outcome of our justificatory investigations, not their premise.

Philosophy
University of Arizona

Notes

[1] These terms, of course, derive from John Rawls. See his *Political Liberalism*, paperback edn. (New York: Columbia University Press, 1996), pp. 48ff.

[2] I intend this as a general concept, which is filled out differently by various public reason liberalisms. Thus what I have called the "Members of the Public" in *The Order of Public Reason* [(Cambridge: Cambridge University Press, 2011), esp. chap. V] is a particular specification of the members of the justificatory public.

[3] This dispute is central to Kevin Vallier's *Liberal Politics and Public Faith: A Philosophical Reconciliation* (PhD Dissertation, University of Arizona, 2011), esp. chaps. 8–10.

[4] Rawls, *Political Liberalism*, p. 81. Rawls identifies three other aspects of this moral sensibility.

[5] Paul Weithman, *Why Political Liberalism? On John Rawls's Political Turn* (New York: Oxford University Press, 2010).

[6] Ibid., p. 59.

[7] Rawls, *Political Liberalism*, p. 392.

[8] Weithman, *Why Political Liberalism?*, p. 54.

[9] Ibid., p. 49.

[10] This, of course, is a deeply ambiguous idea; I hope to clarify it as we proceed.

[11] Gillian K. Hatfield and Stephen Macedo, "Rational Reasonableness: Toward a Positive Theory of Public Reason," *Law and Ethics of Human Rights* (Israel), forthcoming.

[12] Rawls, *Political Liberalism*, p. 38; Weithman, *Why Political Liberalism?*, p. 333.

[13] Rawls, *A Theory of Justice*, revised edn. (Cambridge, MA: Harvard University Press, 1999), p. 505. See Weithman, *Why Political Liberalism?*, pp. 47ff.

[14] Rawls, *A Theory of Justice*, p. 497.

[15] Ibid., pp. 504–5.

[16] Weithman, *Why Political Liberalism?*, p. 53. It may be thought that this simply cannot be the correct analysis of Rawls's public reason liberalism since it is an "ideal" theory that presupposes strict compliance with the principles of justice. However, the fact that

members of the justificatory public endorse principles under the supposition of strict compliance, and that Rawls's proposed institutions are generally supposed to operate under full compliance, do not imply that he believes that actual compliance can simply be postulated, and so the temptation to non-compliance assumed away. Principle *P* may well be the principle that members of the justificatory public would choose supposing full compliance, and qua members of the justificatory public, on the basis of *R* they may be prepared to act on *P*. But if their less restricted set of reasons *U* generally endorses non-compliance, the just society will be unstable.

[17] I am paraphrasing Rawls here. See *Political Liberalism*, p. 386.

[18] On justificatory defeaters see my *Justificatory Liberalism* (New York: Oxford University Press, 1996), p. 66–70.

[19] Rawls, *Justice as Fairness*, p. 89.

[20] We might note here that the problem of ineffective endorsement could be overcome if the typical member of society endorses them on the basis of *U*; as Weithman says of Rawls's account in *Theory*, "congruence need not obtain 'person-by-person'." *Why Political Liberalism?*, p. 59. The crucial thing is that a sufficient body of the citizens affirms *P* on the basis of *U* such that in general endorsement is effective. In contrast, avoiding defeated endorsement looks to require universal endorsement of *P* on the basis of *U*, at least for the idealized deliberative group. I return to these matters below in §2.

[21] Rawls, *A Theory of Justice*, p. 508.

[22] Ibid.

[23] This, I'm afraid, is an absurdly condensed version of the complex argument from chapter nine of *Theory* but it will have to suffice for our purposes.

[24] Rawls, *A Theory of Justice*, p. 505.

[25] Weithman's book masterly analyzes these considerations.

[26] Rawls, *Political Liberalism*, p. xviii (p. xvi of the 1993 edition). Emphasis added.

[27] Rawls, *Justice as Fairness*, p. 20.

[28] Rawls, *Political Liberalism*, pp. 385ff.

[29] Jonathan Quong, *Liberalism Without Perfection* (Oxford: Oxford University Press, 2011), p. 167. Emphasis added.

[30] Rawls, *Justice as Fairness*, p. 34.

[31] Ibid.

[32] Emile Durkheim, *The Division of Labor in Society*, translated by George Simpson (New York: Glencoe, 1964), Book One.

[33] See Weithman, *Why Political Liberalism*?, chap. 10.

[34] *The Order of Public Reason*, chap. 7.

[35] See Rawls, *Political Liberalism*, pp. xlvii–l, 223–27; Rawls, *Justice as Fairness*, pp. 133–4.

[36] Weithman does not see this as a deep problem. *Why Political Liberalism?*, pp. 333ff.

[37] See *The Order of Public Reason*, §19.3.

[38] I am focusing here only on the problem of defeated endorsement.

[39] Here we relax the assumption that *U* contains *R*.

[40] This is my strategy in Chapter V of *The Order of Public Reason*.

[41] On path-dependence, see my *On Philosophy, Politics, and Economics* (Belmont, CA: Thomson-Wadsworth, 2008), pp. 164ff.

[42] Cristina Bicchieri, *The Grammar of Society: The Nature and Dynamics of Social Norms* (Cambridge: Cambridge University Press, 2006), p. 92.

[43] I develop this idea in "The Property Equilibrium in Our Liberal Social Order (Or How to Correct Our Moral Vision)," *Social Philosophy & Policy*, forthcoming. My account draws on Herbert Gintis, *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences* (Princeton: Princeton University Press, 2009), esp. chap. 11.

[44] I believe that Gillian K. Hatfield and Stephen Macedo make this error in their "Rational Reasonableness." Drawing on a model developed by Hatfield and Barry Weingast ["What is Law? A Coordination Model of the Characteristics of Legal Order," *University of Southern California Law School, Law and Economics Working Paper* Series, 2010, paper 123] they argue that successful legal coordination requires public "common logics" that allow us to anticipate each other's reaction to violations (especially punishments). I cannot consider the complexities of the model here, but I note (*i*) the common logics employed in the Hatfield and Weingast model are critical because they yield shared expectations of the behaviors required by the rule and its enforcement, and (*ii*) their model explicitly disallows that one could ever come to infer the "personal"

scripts ("logics") used by others, and so by stipulation only common knowledge of a "common logic" can provide the basis of coordination. Given the severe limits the model's very simple account of coordinated behavior via norms and laws, it is surprising that Hatfield and Macedo (in their "Rational Reasonableness") and Macedo (in "Why Public Reason? Citizens Reasons and the Constitution in the Public Sphere" http://ssrn.com/abstract=1664085) appear to believe the model provides significant support for the necessity of "shared public reasoning." I consider some of the problems with the common knowledge assumption in section 3.2.

[45] Weithman, *Why Political Liberalism?*, p. 49. Emphasis in original.

[46] Ibid., p. 328. Emphasis in original.

[47] Rawls, *Political Liberalism*, p. 387.

[48] For Macedo, see his "Why Public Reason?" and Hatfield and Macedo, "Rational Reasonableness."

[49] Rawls revised his view of to the extent to which this disallows appeal to comprehensive doctrines in political discourse concerning matters of basic justice. See Weithman, *Why Political liberalism?*, pp. 329ff; Rawls, "Public Reason Revisited" in his *The Law of Peoples* (Cambridge, MA: Harvard university Press, 1999), pp. 131-80.

[50] See Rawls, *Justice as Fairness*, p. 89.

[51] See Bicchieri, *The Grammar of Society*, pp. 153-57.

[52] See Ken Binmore, *Natural Justice* (Oxford: Oxford University Press, 2005), p. 68.

[53] In a slightly different game, sometimes called "the Stag Hunt," where one is indifferent between one's unilateral defection and mutual defection, communication that one intended to cooperate would have an unequivocal message.

[54] Weithman, as well as Hatfield and Macedo, endorse the common knowledge assumption. See Weithman, *Why Political Liberalism?*, e.g., p. 328, Hatfield and Macedo, "Rational Reasonableness."

[55] See Gintis, *The Bounds of Reason*, chap. 5.

[56] See Bicchieri, *The Grammar of Society*, p. 11; *The Order of Public Reason*, p. 167.

[57] See *The Order of Public Reason*, pp. 168-72.

[58] See Weithman, *Why Political Liberalism?*, p. 338.

[59] Compare Bicchieri, *The Grammar of Society*, pp. 196ff.

[60] See Brian Skyrms, *The Stag Hunt and the Evolution of Social Structure* (Cambridge: Cambridge University Press, 2004), chap. 3.

[61] The ability to detect cheaters on rules is a basic human proficiency. See *The Order of Public Reason*, §8.

[62] Of course the epidemiological dynamic can go the other way, inducing compliance. I assume that we do not wish stability to depend on a hope for a countervailing tendency.

[63] For evidence, see *The Order of Public Reason*, §7. It should not be thought that punishment has no place in Rawlsian stability arguments; see *A Theory of Justice*, p. 504.

[64] *The Order of Public Reason*, §7. See also my "Retributive Justice and Social Cooperation" in *Retributivism: Essays on Theory and Practice*, edited by Mark D. White. (Oxford: Oxford University Press, 2011), chap. 4.

[65] Rawlsians have begun to perceive the importance of decentralized punishment for norm maintenance. See Hatfield and Macedo, "Rational Reasonableness."