European Journal of Philosophy

DOI: 10.1111/ejop.12020

The Fragility of Consensus: Public Reason, Diversity and Stability

John Thrasher and Kevin Vallier

Abstract: John Rawls's transition from *A Theory of Justice* to *Political Liberalism* was driven by his rejection of *Theory's* account of stability. The key to his later account of stability is the idea of public reason. We see Rawls's account of stability as an attempt to solve a mutual assurance problem. We maintain that Rawls's solution fails because his primary assurance mechanism, in the form of public reason, is fragile. His conception of public reason relies on a condition of consensus that we argue is unrealistic in modern, pluralistic democracies. After rejecting Rawls's conception of public reason, we offer an 'indirect alternative' that we believe is much more robust. We cite experimental evidence to back up this claim.

1. Introduction

John Rawls's early conception of stability required a substantive 'congruence' between the right and the good (Rawls 1999: 496–505).[1] He later rejected this view, however, believing it was untenable in light of the fact of reasonable pluralism. Pervasive disagreement among reasonable members of a well-ordered society will inevitably lead to a breakdown in congruence, destabilizing institutions based on Justice as Fairness.[2] In *Political Liberalism*, Rawls attempted to avoid this problem by developing a conception of stability 'for the right reasons' that obtains so long as citizens come to affirm a political conception of justice consistent with their distinct reasonable comprehensive doctrines.[3]

Public reason, on this account, is the critical mechanism that generates stability. The public, deliberative use of reason and argument drawn from the shared values of a political conception of justice is key to any understanding of public reason. The requirements of public reason are specified by a number of 'guidelines' the most important of which is the duty of civility (*WPL*: 237). By offering and responding only to public reasons, citizens comply with the duty of civility and thereby build social trust. Social trust, in turn, stabilizes the political conception of justice. Citizens come to trust one another and to support that political conception for the right reasons, public reasons related to their own comprehensive doctrines. In this way, civility functions as an *assurance mechanism*, helping citizens to view one another as jointly involved in building a legitimate and just social order. The result, Rawls hoped, would be a well-ordered society of citizens in a dynamically stable equilibrium—whose social dynamics would be self-reinforcing.[4]

European Journal of Philosophy ••:•• *ISSN 0966-8373 pp.* ••–••

We claim that, contrary to Rawls's hopes, the conception of public reason that places the duty of civility at its center is inadequate to maintain a self-reinforcing social equilibrium. Our argument follows Rawls in modeling stability as a kind of mutual assurance game. Nevertheless, contra Rawls and his defenders, we show that the assurance generated by the duty of civility is *fragile*. It can only produce stability under unrealistic assumptions. Furthermore, these assumptions are the very ones that Rawls rejected when he changed his account from *TJ* to *PL*. For Rawlsian public reason to maintain stability, it is necessary to assume away the very diversity that Rawls was rightly so concerned with and, instead, assume that members of the well-ordered society share common knowledge, beliefs, and goals. Without that assumption, stability is susceptible to cascading breakdowns, as we will show below. Public reason as social deliberation, what we call *direct public reason*, can be easily undermined even in a well-ordered society.

In response to the problems with Rawls's conception of public reason, we develop an alternative view on which stability is produced through a process of *indirect public reason*—i.e., without direct, deliberative assurance that all affirm the political conception. To show this, we offer a *correlated equilibrium* model of stability that contrasts with Rawls's Nash Equilibrium model. Rawls's conception of stability as an equilibrium depends on the implausible consensus condition. In contrast, a correlated equilibrium model can dispense with the idea that members of the well-ordered society are in consensus. It can instead generate and maintain social stability in the face of diversity and disagreement through a social mechanism known as a *choreographer*.[5] We argue that well-ordered societies often contain many choreographers who coordinate behavior and generate stable patterns of conduct over time. On this account, stability is provided indirectly when citizens' behavior and diverse modes of dialogue produce convergence on common principles and rules. We believe our model is superior to Rawls's for two reasons: (i) it can generate stability under a wider range of favorable conditions; and (ii) it relies on a more realistic conception of social communication and diversity within liberal democracies.

The next section of this article explains the importance of stability and Rawls's conception of stability in *PL.* Section 3 analyzes *PL*'s account of stability based on direct public reason. In section 4, we argue that Rawls's direct conception of public reason is *fragile* and liable to breakdown in modern liberal societies characterized by reasonable pluralism. Section 5 develops an indirect conception of public reason that can meet the challenges we advanced in section 4. Section 6 concludes by showing how an idea of indirect public reason transforms political liberalism in salutary ways.

2. The Problem of Stability

Contract theorists from Hobbes to Rawls have been concerned with how order is possible in a diverse and contentious social world. If individuals deeply and persistently disagree about life's most fundamental questions, many have

doubted whether it is possible to create and maintain a set of rules that all could endorse in order to resolve disputes and organize cooperation. If it is not possible to find some kind of agreement on a society's basic rules, then such a society's institutions will be unstable. To be stable, a system of social rules must have the continuing assent of the diverse populations typical of modern liberal democracies. Such stability is highly desirable: without it, a conception of justice either cannot perform its central task of creating the conditions for mutual cooperation, or must be imposed by the use of force. Neither option is consistent with the liberal conception of a just society.

Paul Weithman has recently argued that Rawlsian stability is best understood as a 'condition of general equilibrium' where 'everyone knows that everyone else acts justly, and each replies to the justice of others by being just himself' (*WPL*: 44). Stability comes in two types: 'inherent' and 'imposed' (ibid.). Stability is imposed when a society's institutional structure are stabilized through the sovereign state's imposition of political power, independently of (and even despite) the reasons of its citizens. Rawls wished to show that 'inherent stability' is possible (*TJ*: 436). Inherent stability obtains when a 'society that is well-ordered ... generally maintains itself in a just general equilibrium and is capable of righting itself when the equilibrium is disturbed' (*WPL*: 45). 'Forces within the system' must generate this equilibrium (*TJ*: 401). In *PL*, the idea of inherent stability becomes 'stability for the right reasons' (*PL*: xliii).[6] A conception of justice is stable for the right reasons when citizens comply with principles of justice for moral reasons and out of moral motives, not merely from accidental or pragmatic considerations. Inherent stability, or stability for the right reasons, is attractive because it is based on reason, not force. Consequently, inherently stable institutions respect the freedom and equality of citizens.

Let us briefly review Rawls's conception of inherent stability in *TJ*. Part III of *TJ* argues that members of the well-ordered society will come to affirm Justice as Fairness as regulative of their actions. Specifically, they endorse Justice as Fairness because they see it as required by their 'sense of justice.' Nevertheless, Rawls openly asks whether each person has reason to maintain her sense of justice. After all, some will see that they can benefit by suppressing their sense of justice and taking advantage of the same sense of justice in others (*TJ*: 435). This creates an assurance problem, as members of the well-ordered society will realize that such defection is a live option for many people, even for those with a sense of justice because citizens may fear that others, less scrupulous than themselves, will not abide by the rules of justice. In such a situation, acting in accordance with one's sense of justice would open one up to the exploitation and predation of others who are less concerned with acting justly. Justice, on Rawls view, is not meant to be a suicidal enterprise. Citizens must feel assured that if they act justly, others will do the same. This is the heart of the assurance problem for stability. Without assurance, the public conception of justice will not be stable because citizens will, fearing predation, reasonably question whether they should act on their sense of justice.

Rawls attempted to solve this assurance problem by arguing that citizens have independent reason to maintain their sense of justice outside of the benefits it brings them. But they only have such reason when they endorse their sense of justice as part of their good. Rawls calls this process 'congruence', which links 'the concepts of justice and goodness' (*TJ*: 498). Congruence obtains when citizens' sense of justice are endorsed from within their 'thin theory of the good'—their conception of the good separable from their sense of justice. The case that Rawls makes for congruence, however, is conditional. It is only rational to maintain one's sense of justice when others do the same (*WPL*: 67). Rawls writes, 'even with a sense of justice men's compliance with a cooperative venture is predicated on the belief that others will do their part' (*TJ*: 336). Weithman argues that this mutual expectation of compliance leads members of the well-ordered society into a kind of *Nash Equilibrium*, where each citizen has reason to act in a certain way because it is the best response to the actions of everyone else (*WPL*: 64). No citizen can improve her prospects by unilaterally changing her mode of interaction. Rawls employs this language when he writes, 'I should like to show that these principles [principles of justice] are everyone's best reply, so to speak, to the corresponding demands of other' (*TJ*: 103). In the well-ordered society, according to Rawls, individuals lack sufficient incentive to deviate from the principles of justice. In this way, Justice as Fairness acquires inherent stability, as social divergence from Justice as Fairness is self-correcting.

The greatest threat to inherent stability is the breakdown of assurance. For Rawls, stability in a well-ordered society is fixed through a dynamic process of multiple interactions between many persons over time (*TJ*: 434). These citizens must be assured that they all affirm Justice as Fairness. If not, they will rightly fear being taken advantage of and may preemptively opt out of the requirements of justice as a result. Rawls claims that citizens may come to 'lack full confidence in one another' and 'may suspect that some are not doing their part, and so they may be tempted not to do theirs' (*TJ*: 211). This breakdown quickly leads to each individual seeking her own good rather than complying with Justice as Fairness. If this suspicion becomes general, it may 'eventually cause the whole scheme to break down' (ibid.). Thus, for Rawls, these concerns about assurance arise even among members of a well-ordered society with a robust sense of justice.

Despite the limitations of two-person models, we can illustrate Rawls's conception of the assurance problem as a two-person assurance game, with the players representing two randomly selected segments of the population. In this model, symmetrically represented populations have two options. Firstly, they can coordinate on a compliance equilibrium that maximizes mutual benefit, thereby acting on the public principles of justice. This option is only beneficial if each population can assure the other that it will act on the public principles of justice. Otherwise, they can choose the less risky, less beneficial alternative of acting on their thin conception of the good—the non-compliance equilibrium. If they coordinate on principles of justice they will benefit most; however, if one party decides to act from her thin conception of the good, she gains more and

the other person gains less than they would if they had both acted from their (thin) good. This assurance game is represented below:

	Comply	**Don't Comply**
Comply	4,4	2,3
Don't Comply	2,3	3,3

This game has two pure strategy Nash equilibriums: both comply with the principles of justice and both act on their (thin) good. To benefit from acting on the principles of justice, each person must be assured that others will do the same and vice versa. When there is a lack of confidence that the other party will comply, acting on one's (thin) good is the less risky strategy. If we can solve the assurance problem, then the optimal equilibrium can be maintained.

To solve this mutual assurance problem, Rawls postulates that members of the well-ordered society would affirm *ideals* that would give them independent reason to endorse their sense of justice and thus Justice as Fairness.[7] This endorsement makes Justice as Fairness so central to one's good that there is no point in deviating from the public conception of justice. Each person in the well-ordered society will be motivated by a sense of justice to comply with the public conception of justice and will know that everyone else is motivated in the same way. Hence, common knowledge of compliance obtains and mutual assurance is preserved.[8]

Rawls eventually came to believe, however, that *TJ's* idea of the well-ordered society was 'unrealistic' because it assumed that citizens shared the ideals associated with Justice as Fairness.[9] In order to share these ideals, all members have to accept a conception of the person that includes a desire 'to be and to be recognized by others as being a certain kind of person' (Rawls 1974: 12–13). If members of the well-ordered society reject this conception and the partial 'comprehensive philosophical doctrine' associated with it, inherent stability will break down (*PL*: xviii). This is why Rawls's concern with reasonable pluralism is so important. It is due to reasonable pluralism that citizens will reject the ideals that buttress compliance with Justice as Fairness. We can see, then, that *TJ*'s well-ordered society contains an internal dynamic that destroys the mutual assurance mechanism intended to produce inherent stability. Reasonable pluralism will lead citizens to affirm different ideals and so their independent reason to comply with Justice as Fairness may wane through entirely natural and admirable developments in citizens' philosophical, social and religious commitments.

In response to this problem, Rawls converted Justice as Fairness into a 'political conception of justice' in order to relax the degree to which citizens must share certain ideals (*PL*: 134). The political conception in turn becomes the object of an 'overlapping consensus' where citizens endorse the political conception from within their distinct but reasonable comprehensive doctrines. So citizens can produce stability without endorsing the partial comprehensive doctrine associated with Justice as Fairness in *TJ*. Rawls also relaxed the congruence condition, arguing that stability is possible so long as each reasonable view is

'either congruent with, or supportive of, or else not in conflict with, the values appropriate to the special domain of the political as specified by a political conception of justice' (*PL*: 169). In *PL*, so long as citizens have sufficient reason to endorse the political conception *regardless* of their reasonable comprehensive doctrines, stability is within reach. Reasons provided by the political conception need merely 'normally outweigh' values and principles that might contradict or undermine them (*PL*: 156).

Yet, stability requires assurance even in *PL* since citizens may still doubt that others will abide by and internalize the political conception. Rawls must still claim that citizens endorse the political conception based on shared ideals that Rawls came to call 'political' ideals (*WPL*: 271). If members of the well-ordered society find the ideals associated with the political conception inspiring for their own sake, then they have reason to internalize Justice as Fairness even if they think others may not do the same. They will thereby have reasons to internalize Justice as Fairness besides those considerations that favor the conditional endorsement of accepting a common conception of justice.

Rawls defends the need for shared ideals by arguing that 'a political conception assumes a wide role as part of public culture' which contains a 'certain conception of citizens as free and equal' (*PL*: 71). The political conception includes an 'ideal of citizenship' learned under conditions of 'full publicity' (ibid.). Citizens who see themselves as free and equal wish to live under principles that each person has reason to accept. Rawls therefore emphasized the ideal of public reason because citizens can use public reasons to justify political principles to one another. These reasons develop and refine an overlapping consensus, which in turn produces and preserves inherent stability.[10] Note the continuing presence of the assurance problem: despite the existence of an overlapping consensus, citizens of the well-ordered society must still assure one another that it exists.

If Rawls can resolve the assurance problem, he can establish that citizens have overriding reason to maintain their desires to live up to the ideals of Justice as Fairness (*WPL*: 299). However, solving the assurance problem requires that everyone know that everyone else also accepts the public conception. Even the suspicion that principles of justice are not shared can potentially undermine assurance. To quiet this concern, Rawls argued that our conception of ourselves as free and equal implies that we will engage in political justification governed by *the liberal principle of legitimacy*. This principle holds that political power is justified only when it is compatible with a constitution that all can 'endorse in light of principles and ideals acceptable to their common human reason' (*PL*: 137). If citizens comply with the principle, then assurance is easier to generate as political justification occurs in commonly accepted terms. In this way, citizens can tell that one another affirm the political conception. Nevertheless, even the liberal principle of legitimacy requires refinement by employing 'guidelines and criteria' that are one of two 'companion parts of one agreement' made in the original position (*PL*: 226). With guidelines selected, reasonable comprehensive doctrines will undergo a 'transformative effect' where they become more compatible with the political conception. This, in turn, leads to inherent stability.

In sum, an internal dynamic created by reasonable pluralism leads to Rawls's mature conception of stability. Citizens of the well-ordered society must converge on a political conception of justice that they have reason to endorse because of the ideals of citizenship, friendship, etc. that it contains (see *WPL*: 270–300, esp. 287–95). In this way, citizens can maintain mutual assurance of the public conception of justice. It is crucial that each person know that most others also endorse the political conception. Consequently, they must be disposed to speak to one another in terms of public reason in the public sphere. Even many followers of Rawls miss these subtleties. For instance, Martha Nussbaum considerably oversimplifies matters when she maintains, 'political liberalism can provide stability, provided that the holders of various comprehensive doctrines care sufficiently about respect for persons' (Nussbaum 2011a: 42). It is not enough for members of the well-ordered society to care about respecting persons. Rawls's conception of stability needs assurance-generating mechanisms even among persons that see each other as free and equal. We will now analyze Rawls's preferred mechanism for maintaining assurance: public reason.

3. Public Reason and the Duty of Civility

Public reason is, at least partly, a mechanism of mutual assurance. When citizens reason in a public fashion, they can provide the assurance necessary to give each other reason to endorse the political conception. Thus, if citizens use public reasons to justify the use of political power, they also assure one another that they are committed to the public conception of justice. The offering of public reasons, on Rawls's view, displays a commitment to the political conception because citizens primarily speak in terms of political values and reasons rather than reasons that derive from their comprehensive doctrines. Further, if citizens speak in political terms, it suggests that they at least partly affirm the political conception, given their willingness to discuss political issues within its confines. In this way, the public use of reason can provide evidence for the existence of an overlapping consensus (*WPL*: 328). For Rawls, doubts are 'put to rest' when 'leaders of the opposing groups ... present in the public forum how their comprehensive doctrines do indeed affirm [the] values [of the political conception]' (*PL*: 249). It is important to emphasize the importance of public reason in light of the fact of reasonable pluralism. Since reasonable disagreement is inevitable, citizens must justify the political conception to themselves in their own terms. As Rawls claims, this 'full justification' is 'left to each person' (*PL*: 386). Given our diverse affirmations, we have strong reason to assure one another of our allegiance to the political conception. We must be disposed to speak in agreed upon and commonly accepted terms to produce assurance (Rawls 1997: 786).

Citizens must also offer public reasons in accord with what Rawls calls '*the duty of civility*'. The duty of civility is a moral duty to provide public reasons in a public forum and to restrain one's use of comprehensive reasons when

constitutional essentials are at stake. This duty performs an epistemic function because it is a guideline for the use of public reasons.[11] Specifying this function is a subtle matter. Rawls changed the content of this duty not once, but twice. He initially entertained the 'exclusive view' of public reason that 'on fundamental political matters, reasons given explicitly in terms of comprehensive doctrines are never to be introduced into public reason' (*PL*: 247). The nice feature of the exclusive view is that it prevents the 'noise' associated with comprehensive doctrines from undermining mutual assurance. Rawls, though, quickly backed off the exclusive view and moved to the 'inclusive view' of public reason which permits 'citizens . . . to present what they regard as the basis of political values rooted in their comprehensive doctrine, provided they do this in ways that strengthen the idea of public reason itself' (ibid.). He later rejected even the inclusive view for the 'wide view' of public reason which holds that citizens can introduce comprehensive reasons into political discussion so long as 'in due course proper political reasons . . . are presented that are sufficient to support whatever the comprehensive doctrines introduced are said to support' (Rawls 1997: 784). By complying with the wide view, members will provide one another with explanations for why they comply with the political conception. This process of providing explanations to one another generates assurance. Rawls claims that the advantage of 'the mutual knowledge of citizens' recognizing one another's reasonable comprehensive doctrines bring out a positive ground for introducing such doctrines' (Rawls 1997: 784).

One reason that Rawls may have embraced the wide view is that it permits diverse procedures for providing assurance. For instance, he maintains that the wide view permits 'reasoning by conjecture' whereby citizens with one reasonable comprehensive doctrine can engage those with distinct and conflicting reasonable comprehensive doctrines on the latter doctrines' own terms (Rawls 1997: 783). In this way, Rawls permits citizens to assure one another even with the use of non-public reasons. Further, the duty of civility is plausibly read as requiring relatively little effort on the part of citizens, giving them less disincentive to speak in public terms. According to Paul Weithman, this interpretation allows Rawlsian citizens to use their comprehensive doctrines 'without adducing public reasons in support of their positions, so long as their doing so does not lead others to doubt that they acknowledge the authority of the public conception of justice' (*WPL*: 330). If doubts do not arise, the proviso is *never triggered*. Citizens need only comply with the wide view if they think 'assurance is actually needed' (*WPL*: 331).

Publicly reasoning in accord with the duty of civility consistently reinforces assurance in a well-ordered society. This creates a condition of common knowledge or *consensus*.[12] This condition, according to Rawls, builds over time. As Weithman argues, 'common knowledge of an overlapping consensus is not based only, or even primarily, on what citizens say issue-by-issue in the public forum' but rather that such knowledge builds up 'over time' (ibid.). This explains why Rawls claimed 'the details about how to satisfy this proviso [the wide view] must be worked out in practice and cannot feasibly be governed by

a clear family of rules given in advance' (Rawls 1997: 784). We cannot make sense of how to satisfy the proviso in the abstract. Thus, the duty of civility maintains assurance over an extended series of iterated interactions among millions of citizens, many of which do not trigger the duty at all.[13]

Using the duty of civility within public reason to ensure stability is an elegant solution to the mutual assurance problem. Nevertheless, as we argue in the next section, it is not robust enough to maintain assurance over time. Consensus based stability is unlikely to survive within diverse modern liberal democracies.

4. The Fragility of Consensus

Stability is conditional on the maintenance of mutual assurance in the well-ordered society. Mutual assurance is based on the common knowledge that members of the well-ordered society are committed to abiding by the requirements of the public conception of justice—on consensus. In this section, we reject the consensus requirement because it is difficult to maintain and because it can be easily undermined. In short, it is fragile. Much of our argument is drawn from experimental and behavioral data that shows the fragility of maintaining consensus in laboratory settings (Gintis 2009: 153). This evidence is strong, we argue, because if it is difficult to maintain stable equilibria via consensus in controlled laboratory environments, it should be even harder to maintain them in the diverse and often contentious environment of modern liberal democracies. Consequently, we argue that consensus based stability is fragile and liable to disintegration even under favorable conditions. We describe Rawls's model of providing consensus as a conception of *direct public reason*, where citizens' direct deliberation with one another provides assurance.[14] However, as we will argue below, assurance provided by direct public reason can be undermined by noise and drift.[15] Direct public reason is thereby unable to maintain stability. To have a stable system of justice without rejecting wide conception of the duty of civility, we must move to an indirect conception of public reason.

We stress here that the problems we raise are based on game theoretic and experimental models that should apply even to members of the well-ordered society with a robust sense of justice. Thus, assurance problems plague Rawls's view they can do so even under idealized conditions. We do not claim here that Rawls's model is unrealistic so much as that the model cannot generate Rawls's intended result.

Before we develop our criticisms, we should say a bit more about the sense in which Rawls advocates a direct model of assurance provision. It is true that Rawls does not rely exclusively on a direct method of assurance, given that laws can often provide assurance. Consider: '[I]n a well-ordered society there would be no need for the penal law except insofar as the assurance problem made it necessary' (*TJ*: 277). Nevertheless, law alone cannot produce stability for the right reasons; it can only keep people in line. One might counter that Rawls can

appeal to another alternative mechanism: observation of mutual compliance with the political conception. Observation without direct assurance, however, is subject to different interpretations; citizens might think that others are simply complying due to social or political pressure or because they are confused. Unless citizens are prepared to directly assure one another, simple observation cannot do the job. We believe Weithman agrees, since citizens must provide assurance 'by actually adopting and reasoning from the "unified perspective" the public conception of justice provides'; similarly, the duty of civility 'requires citizens to adopt and deliberate in their "common point of view"' (*WPL*: 331). Therefore, we think we are on solid ground in holding that Rawls's main assurance mechanism is direct. Without direct assurance, the model collapses. Other forms of assurance are insufficient on their own. Rawls arguably thought as much, as evidenced by the amount of time he spent developing the duty of civility vis-à-vis his fleeting remarks about alternatives.[16]

Two problems undermine stability under direct public reason: noise and amplification. Noise is the problem of distinguishing between communication by citizens that signal allegiance to the public conception, and hence assurance, and forms of communication that do not. For instance, once the wide view of public reason is adopted, individuals may present sectarian or self-interested reasons in the public sphere so long as some public justification can be given in due course. Once those other reasons are allowed, however, it will be difficult if not impossible to distinguish public reasons based on the public conception from those that are not so based.

Since citizens are able, on the wide view of civility under direct public reason, to introduce some private reasons into the public sphere, it becomes difficult to tell when citizens are advancing public reasons or not. Specifically, the presence of other forms of reasons may make it difficult to tell when citizens are advancing public reasons in the cacophony of the public sphere. Since the public provision of reasons is the primary way that citizens signal their allegiance to the political conception of justice, true signals of allegiance must be easily distinguishable from noise. However, on the wide view, deliberative noise would likely be substantial. This holds even if everyone sincerely attempts to signal his or her allegiance along with other forms of political discourse.

Consider an example most familiar to American readers.[17] Once sectarian and other reasons are allowed into the public sphere, it is impossible to know, for instance, whether defenders of various restrictions on abortion are presenting reasons that are consistent with the public conception of justice. Are they merely couched in the language of sectarian comprehensive justifications (e.g., religious or moral) or do they reflect a more fundamental rejection of the public conception of justice? It is hard to know. This is the sense in which assurance signals become noisy when sectarian and comprehensive reasons are allowed on the wide view. Noise happens when additional signals are introduced that make it impossible for citizens to distinguish between public communication that provides assurance and public communication that signals a move away from the public conception.

If we follow Rawls in modeling stability as an assurance game, noise can be modeled as 'cheap talk'.[18] Cheap talk is costless or very inexpensive, non-binding communication in a game. Robert Aumann and Sergiu Hart describe cheap talk among players in a game as follows: '[T]he players don't strike, don't get educated, and don't issue guarantees; they simply talk. They may or may not tell the truth, and may or may not believe each other' (Aumann and Hart 2003: 1619). The problem is not that members of the well-ordered society will lie to one another about their allegiance to the political conception; only that in noisy conditions they will lose their confidence (i.e., assurance) that they have reason to trust that others will abide by the political conception. In turn, this will tend to undermine consensus. Traditional game theory clearly suggests this conclusion. Robert Aumann (1990), for instance, argued that in an assurance game similar to the one we have been discussing, the mutual assurance equilibrium can not be sustained under conditions of cheap talk, where there are gains for moving to the risk-dominant equilibrium. In the understanding of the stability model that Rawls uses we would, therefore, have good reason to think that noise would undermine consensus based stability when that consensus is based on noisy assurance signals.[19]

We might question how troubling this conclusion should be. Do we really have good reason to believe that members of the well-ordered society will do what traditional game theory suggests and move away from the consensus when consensus is based on reliable signals of assurance?[20] One might argue, for instance, that since 'talk is cheap' it should have little effect on the stability of a given equilibrium. A so-called 'babbling equilibrium' might not necessarily differ from a mutual assurance consensus equilibrium. Brian Skyrms (2004: 65–82) has argued, however, that this is a mistake. In large populations with random encounters, Skyrms shows that the introduction of cheap talk can destabilize a mutual assurance consensus equilibrium.

The experimental evidence comes to similar conclusions, though the evidence is slightly weaker.[21] In general, pre-play, non-binding communication does seem to increase the tendency to cooperate in social dilemmas (Sally 1995). This tendency decreases significantly, however, as the numbers of interaction increases (Sally 1995: 78). Furthermore, Cristina Bicchieri (2002) found that the tendency to cooperate in these cases significantly increases if some form of group identity is created. There are other experimental settings, however, where cheap talk does regularly undermine assurance. Subjects will initially cooperate with others, but once even a small number of players decide that they have reason to pursue their own gain at the expense of the consensus, cooperation rapidly breaks down. Surprisingly, in these same experiments, cooperation *decreases* when cheap talk is permitted before each round of game play as a method of assuaging worried defectors (Wilson and Sell 1997: 695).

In one experiment, researchers found that cheap talk tended to be more effective in a face-to-face setting. This led them to conclude that the effectiveness of cheap talk communication crucially depends upon 'institutional context in which that communication takes place' (Wilson and Sell 1997: 714). Public

forums where citizens interact on a personal basis make signaling loyalty easier relative to larger, impersonal settings. Of course, in a well-ordered society interactions will be largely anonymous and vast in scale. Consequently, the assurance situation that will obtain in the well-ordered society will probably resemble Skyrms's model noted earlier. We have good reason to believe then, that the effect of noise will tend to undermine stability, though it is impossible to identify the degree to which it will do so in the abstract. There is good reason to believe then that, given the formal and experimental evidence, that cheap talk and noise will lead consensus-based equilibriums maintained by direct public reason on a large scale to become unstable.

Rawlsians are bound to claim that much of this experimental work is irrelevant for criticizing Rawls's view because real-world subjects differ substantially from members of the well-ordered society. Rawls, for instance, introduces the notion of an *ideal* of public reason that could, potentially, be used to sanction legislators and other public figures that introduce noisy or non-public reasons (Rawls 1997: 767–9). Consequently, Rawlsians will argue that we have little reason to think that these real-world problems will arise in Rawls's model. The problems we cite, however, do not depend on motivations other than those that derive from the rational and reasonable psychologies of members of the well-ordered society. The empirical and formal literature suggests that cheap talk will pose a problem even if all parties are motivated by their good and their sense of justice, as they may still have doubts that others' words will not be backed up by actions. After all, even Rawls thinks members of the well-ordered society have reason to defect if they do not believe others will do their part. The reason for this is that Rawls allowed comprehensive reasons that support the political conception to enter into public discourse, so it will be hard to sanction public officials for introducing non-public, and so often noisy, reasons, either by using the ideal of public reason or in the background culture.

The second problem for consensus on the wide view of direct public reason is an amplification of noise. Amplification occurs when errors in communication multiply over large numbers of interactions, a phenomena we term 'informational drift'. Small errors, when so multiplied, can quickly lead to a cascade of misunderstanding, miscommunication, and divergence in interpretation. This informational drift can create 'informational cascades' that can dramatically undermine mutual assurance consensus equilibria.[22] Lisa Anderson and Charles Holt explain that an informational cascade 'occurs when initial decisions coincide in a way that it is optimal for each of the subsequent individuals to ignore his or her private signals and follow the established pattern' (Anderson and Holt 1997: 847)—i.e., a signaling error can become a public norm. A classic example is a financial market bubble. Vernon Smith and David Porter created small-scale asset bubbles in laboratories where traders continued to sell and buy assets even after the underlying value went to zero (Porter and Smith 1994). Furthermore, continued experience in 'bubble' markets does not significantly reduce the prevalence of bubbles in these conditions (Smith et al. 2008). Nobody wanted to be the last holding the asset, and therefore, many continued to sell zero-valued

assets up to the point of market collapse. Robust bubbles formed *especially* when public information was 'cheap' and noisy (Porter and Smith 1994: 122). This result is robust across many different population groups, leading us to believe that it is not a product of a particular psychological assumption.

For similar reasons, we sometimes see a similar effect in stock market panics. There is confusion about the underlying asset value of certain stocks leading traders to interpret the behavior of other traders as signaling information about underlying assets. The problem is that, in circumstances of noisy information, signals can be misinterpreted. The fact that one of these cascades catches on does not necessarily communicate any information about the underlying asset itself, but instead merely reflects the incorrect beliefs of the traders. Cascades can spread quickly and 'can be upset by the arrival of new public information' (ibid.). As a result, cascades are frequently unpredictable. In controlled laboratory settings, experimenters find that agents are most susceptible to information cascades when they are asked to make public decisions based on private information.

This susceptibility is reflected in the stability problem faced by members of the well-ordered society. They are more like the subjects in the asset bubble experiments or traders in financial markets than traders in more stable market situations. Citizens in the public forum use their private information to send public signals about the content of the political conception and the guidelines of public reason, along with their respective endorsement of both. Other citizens must interpret those signals as public information often without knowing the source of the signals or their causal history. This is especially true of public figures like members of the United States Congress, where support for a piece of legislation may have less to do with the considerations of justice than with its value for the political actor in terms of future votes or allegiance to ideology. When a public person gives reasons in the legislature, how are we to understand that signal?

Stability maintained by direct public reason is fragile because the wide view of civility permits noisy signaling that can be amplified by informational drift. These problems will arise even in a well-ordered society, as the phenomena do not require that agents be unreasonable or substantially misinformed. Accordingly, it seems that Rawls's attempt to identify a social mechanism to produce mutual assurance fails. One might worry that this failure threatens his entire theoretical project, but to conclude as much would be premature. We believe that by replacing Rawls's conception of direct public reason with a conception of indirect public reason, we can save stability for the right reasons. If we are successful, Rawlsians will have strong reason to endorse an indirect account of public reason.

5. Indirect Public Reason

In the last section, we argued that direct public reason is fragile. The wide view of civility allows costless signaling of reasons, which undermines stability. To

solve this problem, it may appear that we must increase the cost of signaling so that citizens can be confidently assured that their fellows share the public conception of justice. If we increase the cost of noise and informational error, it seems that we must also increase the cost of signaling assurance. Increasing the cost of signaling, however, requires retreating from the wide view to the inclusive or exclusive view of civility, as public discourse would need to be stripped of extraneous talk. Such restrictions should be unattractive to political liberals who care about protecting freedom of speech and expression. Further, the problems of confusion and drift remain, as members of the well-ordered society will be subject to error even under more restrictive conditions.

In response, we might decrease the cost of assurance by permitting more assurance-providing techniques, but doing so seems to open the door to error as cheap talk will become an increasingly common phenomenon. Further, given an increase in methods and a decrease in restrictions, noise will increase as well. Either way, the direct approach to public reason seems like a dead end.

On analogy, we can imagine an assurance game with drivers at an intersection. Assurance mechanisms are required to allow drivers to cross the intersection safely without crashing into one another. First, imagine a case with no stop signs or lights. In this case, drivers have to use their own lights, horns and eye contact to coordinate. This is costly and subject to error, frustration and confusion. Drivers will be forced to slow down considerably or to crash into one another. This is analogous to the conception of direct public reason. One could improve matters with stop signs, say, at a four-way stop. However, as the number of cars increase, the amount of direct coordination required becomes more costly as drivers must still determine who first arrived at the light. We have all been in these situations. Four-way stops are terribly inefficient in high traffic areas. We can see then that all of these methods of direct communication with drivers have serious limitations with respect to the cost of assurance.

The obvious solution to our traffic problem is to install a traffic light. Traffic lights correlate coordination among drivers to an independent, public signal. By following a traffic light, drivers no longer need to directly assure one another of their intentions by signaling. Instead, they realize that it is in everyone's best interest to follow the public signal. The traffic light thereby dramatically reduces the epistemic problems associated with traffic intersections. To put it in game-theoretic terms, the traffic light creates a *correlated equilibrium.* Herbert Gintis calls correlating mechanisms like traffic lights choreographers. On this model, the choreographer generates assurance indirectly. Choreographers significantly aid the process of forming and maintaining assurance because players no longer need to coordinate directly; they gain from simply abiding by the directions of the choreographer (Gintis 2010: 252). With the correlation mechanism in place, no driver can do better by unilaterally deviating from the direction of the traffic light; hence, the new correlating mechanism creates a new, stable convention.[23]

Choreography does not arise *ex nihilo.* For a correlated equilibrium to be effective, the choreographer must use comprehensible signals that parties can respond to effectively. One way this process occurs is through a 'salience'

creation mechanism. Thomas Schelling (1960), in his classic *The Strategy of Conflict*, introduced the idea of salience via his theory of focal points in coordination games. In a pure coordination game, players gain from choosing identical or related strategies. For instance, in a meeting game where two players try to locate one another, each gains only if they both choose to meet at the same spot. In most games, the destination is irrelevant so long as both parties end up there. Schelling (1960: 57–8) pointed out that doing well in such a game requires stumbling on some salient property of one option that the other player will also pick-up on. However, Schelling emphasizes that finding salience in a coordination game is more art than science. Correlated equilibriums will be effective only insofar as parties notice some antecedently salient feature of the world. Conventions often evolve that are salient to many agents, such as basic rules of the road.[24] However, without those antecedent conventions, a choreographer like the traffic light will lack the salience necessary to effectively correlate agents' behavior.

Drift and noise, of course, can occur even with a choreographer. The problem of stability, however, is not that agents cannot deviate from public standards. Instead, the problem is to sufficiently reduce drift and noise so that stability for the right reasons can form and survive. A correlated equilibrium conception of stability under indirect public reason makes the formation problem less severe because the assurance of public endorsement is generated indirectly through a public, external event rather than through noisy, misleading direct assurance mechanisms. The primary challenge for an indirect public reason conception view is to identify the relevant choreographers and to explain how they gain salience within a liberal society. In the next section, we attempt to meet this challenge.

6. Stability in a Liberal Society

Rawls's conception of stability is based on a direct model of public reason, where citizens directly assure one another of their loyalty to the political conception of justice (*WPL*: 331) By directly assuring one another in accord with the duty of civility, citizens of the well-ordered society will comply with the dictates of the political conception. However, if we replace Rawls's direct conception of public reason with an indirect conception of public reason, members of the well-ordered society need only follow the relevant public choreographer and believe that others will do the same. Far less direct assurance is required. On both views, citizens of the well-ordered society must judge that their balance of reasons favors maintaining their desire to follow the political conception. But, the indirect conception of public reason employs assurance mechanisms other than the duty of civility. Because of this, the duty of civility is less important on an indirect conception of public reason. Members of the well-ordered society can rationally accord priority to the political conception so long as their society contains salient publicly recognized choreographers directing them to do so. The

Rawlsian project only needs public choreographers to provide assurance on the indirect model.

Who are the public choreographers in a well-ordered society? At the most general level, choreographers are events, like the changing of the traffic light. However, choreographers can come in a variety of types: (i) deliberate and spontaneous, (ii) individual and group, (iii) macro and micro. Firstly, choreographed events can be produced deliberately or spontaneously. For instance, traffic lights produce choreographed events because civil engineers program them to do so. In contrast, correlated norms can arise spontaneously so long as there is a publicly recognized event that organizes behavior. Secondly, choreography can be performed either by individuals or by groups.[25] A judge may serve as a correlated equilibrium when she issues a verdict, but a congressional body may do the same. Finally, choreographers can exist at different levels of social organization. Assurance is not provided directly, that is between groups of citizens, but instead indirectly through, e.g., recognized fealty to public courts of law. Such recognition allows the courts to play a choreographing role.[26] Alternatively, choreography can arise from a plurality of local choreographers, such as local statutes and ordinances set by citizens and counties. Society-wide stability can be generated at the federal level or in a piecemeal fashion from these local bodies of law.

Choreographers, it seems, are ubiquitous.[27] When we recognize their legitimacy—i.e., when they are salient—they produce inherent stability. Public discussion is only part of this process, as most assurance is *indirect*.[28] On the indirect view, citizens' activities produce stability, but not through direct, deliberative, intentional assurance. For instance, when we argue about the proper interpretation of the First Amendment to the American Constitution, even if we appeal to non-public reasons, we implicitly assume that each party to the discussion affirms the principles underlying it. Our discourse indicates that we all endorse the constitution despite our differing interpretations. In this way, public discourse, even disagreement, produces assurance by indirectly providing evidence of acceptance and compliance with the public conception.[29]

In our view, public choreographers are primarily bodies of norms, often legal though sometimes informal or formal moral norms. By focusing on indirect rather than direct assurance through public reason, the theoretical focus shifts from public reasons to public *rules*. Rawls wants to build discourse around public reasons in order to produce direct assurance. However, if public reasons by themselves cannot generate and maintain inherent stability, then a core motivation for focusing on public reasons is undermined. Shared discourse can help demonstrate our commitment to compliance, but this demonstration can occur in many, often non-public ways. The idea of a correlated equilibrium reorients us towards focusing on public events that consist in the creation, affirmation, revision or rejection of publicly recognizes rules of conduct. The idea of public rules can be understood as a form of indirect public reason.

A key difference in our model of stability is that, in contrast to Rawls, we need not know one another's reasons for complying with choreographers. Rawls

seems open to this when he notes that: '[Members of the well-ordered society] take into account and give some weight to only the fact—the existence—of the reasonable overlapping consensus itself' rather than looking to the comprehensive reasons that citizens affirm the political conception (*PL*: 387). Rawlsians, however, will surely object. Rawls's concerns about publicity indicate that he wants to base social institutions not merely on public rules but on public rationales. Thus, Rawls insists that the 'full justification of the public conception of justice' should be 'publicly known, or better, at least to be publicly available' (*PL*: 67). Unless the rationales are public, something important may be lost, such as a sense of fellow-feeling among citizens. Yet, while *something* important may be lost, we cannot see any reason to think that stability will be.

There may well be some cost to abandoning the publicity of rationales, but given the importance of an adequate account of stability, the costs are likely worth paying. As we have said, on the indirect view, stability does not require the publicity of rationales. For one thing, the correlated equilibrium model shows that inherent stability can be achieved without them, and secondly, making rationales reliably public in Rawls's sense is difficult, given the phenomena of noise and drift. Thus, traditional Rawlsians adopt a faulty conception of stability. Stability for the right reasons can be maintained so long as social processes and institutions associated with the political conception are publicly recognized and followed.

Our model compensates for weaknesses in Rawls's view, but one advantage of Rawls's mature conception of stability is that may explain how a society can move from a modus vivendi to an overlapping consensus with inherent stability. If we can converge on an overlapping consensus, then citizens can directly assure one another of their loyalty to it and even their commitment to creating it. Direct assurance mechanisms clarify how persons can rationally recognize the overlapping consensus and demonstrate their allegiance to it. Indirect assurance cannot get off the ground unless choreographers already exist and their antecedent conditions are already met. How then do we select public choreographers? We cannot appeal to direct assurance for the reasons discussed above. We seem stuck in a chicken-and-egg problem—we cannot comply with choreographers until they are recognized, but we cannot use choreographers to converge on them in the first place.

In response, consider an analogy drawn from ship construction. We can conceive of ship construction in two ways. Ordinarily, a group of builders gets together and produces a blueprint. Once the ship is complete, all recognize that the ship matches the blueprint and can then use the ship to cooperate with one another when they set sail. Rawls's mature conception of stability is akin to construction of this sort. In the first stage, members of the well-ordered society agree on a social 'blueprint' that will satisfy their common aims (the political conception) and then they must check to see if they agree with the final product (the production of an overlapping consensus). Finally, when the ship is finished, they must publicly agree that the ship has been properly constructed in accord with the blueprint (the assurance process).

Contrast the blueprint model with repairs to a ship at sea, following Otto Neurath's famous example. In this case, the ship's crew cannot build a new ship from scratch since they are in the middle of the ocean. But even without a blueprint, they can identify defects, leaks and structural problems. Further, sailors face a fundamentally different problem from ship builders. Their problem is not to discern what a perfect ship looks like, but how to make reasonable, piecemeal improvements. Even if the ship was poorly constructed at first, sailors can construct a better ship over time. The crew need only work off a publicly recognized framework of rules of repair and the use of shared materials.

Our model of inherent stability is closer to the latter form of ship construction, as it assumes that the relevant type of stability is already in place. Some social structure must already be in operation for an overlapping consensus to develop. Thus, any society that develops an overlapping consensus must already have a network of choreographers in operation. These choreographers need not be perfect or even wholly legitimate. Instead, they provide the cooperative 'social capital' out of which a just society can be built. Citizens of a not-quite-well-ordered society can move towards justice with a set of norms and choreographers already active. Note that they can do so only if some of their choreographers are regarded as legitimate. Otherwise, there are no legitimate choreographers available to move society towards a more justified set of norms. Thus, stability for the right reasons is no mere output of the process of political justification but an *input* as well. The indirect idea of public reason holds that we must begin moving towards a just society with imperfect forms of stability that are improved over time. In this way, we can locate a middle ground between merely practical and inherent stability, for most stable liberal democracies combine the two. Sometimes stability is imposed through state violence and oppression by some social classes.[30] Other times, stability is maintained and perfected through a public recognition of fairness.

The Rawlsian ideal of stability becomes a regulative ideal for repairing and refining one's institutions, for moving from the impure social ore of the present to a more just social product. In this way, our model differs from Rawls's. We might interpret Rawls's project in *PL* as a 'possibility proof' for an ideally just liberal democracy. In other words, he merely attempts to specify the conditions under which such a society is possible. Our model is both more and less ambitious. It is more ambitious because we want to show how stability can be reached from our present conditions.[31] It is also less ambitious because it relies more on actual social processes and structures to carry out the process of justification rather than the more deliberative process Rawls envisions.[32]

John Thrasher
Department of Philosophy
University of Arizona
USA
jthrashe@email.arizona.edu

Kevin Vallier
Department of Philosophy
Bowling Green State University
USA
kevinvallier@gmail.com

NOTES

[1] Rawls (1999) is hereafter *TJ*. Many readers may be unfamiliar with Rawlsian 'congruence', but the concept is pivotal. See Weithman (2010; hereafter *WPL*). Throughout, we use *PL* to refer to Rawls (2005).

[2] Some Rawls scholars seem to believe that Rawls's concern with stability was confined to *PL*. Martha Nussbaum (2011b: 2) has claimed that 'the central problem of [*PL*], as contrasted with [*TJ*], is that of stability'. Following Weithman, we think stability is central to *TJ*, though it is mostly discussed in the oft-ignored Part III of *TJ*.

[3] That is, affirmed as *legitimate*, not necessarily as just (*PL*: xlvi).

[4] We follow Weithman in understanding 'inherent stability' and 'stability for the right reasons' as a kind of Nash Equilibrium. In a game, strategies are in a Nash equilibrium when the strategy is the best response to the best response of all the other players. In short, no one can do better by unilaterally changing their strategy to some other strategy. A system that is in equilibrium in this way is inherently stable in just the way that Rawls is concerned with; e.g., Rawls discuses stability as being a condition of 'each man's best reply' to others (*TJ*: §76, 435).

[5] We take this terminology from Gintis (2009: 44).

[6] Rawls dropped talk of 'inherent' stability in *PL*. We use 'stability for the right reasons' and 'inherent stability' interchangeably as the element of 'inherence' we stress is more general than Rawls's.

[7] Weithman discusses these ideals at *WPL*: 81. See also *TJ*: 397–449 for an extensive discussion of how this mechanism works.

[8] We do not claim that common knowledge is required to maintain an assurance equilibrium, only that Rawls believes it is necessary.

[9] Specifically, Rawls thought that the argument of *TJ*: §86 failed. See *WPL*: 234–69 for an extensive discussion of this pivotal shift.

[10] Much molding of comprehensive doctrines to fit the political conception is done privately, among adherents. Thus, public reason is not the only way to create an overlapping consensus. We thank Paul Weithman for this point.

[11] The duty of civility is not meant *merely* as an assurance mechanism. It also helps characterize Rawls's conception of civic culture.

[12] We understand the common knowledge or consensus requirement as holding that every member of the well-ordered society has, in Bayesian terms, 'common priors', including the knowledge of the rules of the game and the rational strategies of each other player. Everyone must also know that everyone else has the relevant knowledge and so on. Arguably, Rawls affirmed a common knowledge requirement, though we deny that a common knowledge requirement is necessary to provide assurance.

[13] If assurance is dynamic, we can see why the duty of civility largely restricts citizens' behavior rather than encouraging it. Since assurance keeps a society in equilibrium as learning builds, the duty need only provide resistance when society seems to move out of equilibrium.

[14] Rawls does not rely exclusively on a direct method of assurance, given that laws can often play this role. Consider: '[I]n a well-ordered society there would be no need for the penal law except insofar as the assurance problem made it necessary' (*TJ*: 277). Nevertheless, the law alone cannot produce stability for the right reasons; it can only keep people in line. Therefore, Rawls's main assurance mechanism is direct. If the direct assurance mechanism fails, given the amount of time Rawls spent developing it, the secondary indirect mechanisms will probably not be strong enough to salvage the system. We thank an anonymous referee for bringing this to our attention.

[15] It is worth pointing out that we, like Rawls, are not really concerned with how the assurance equilibrium is attained; we are only concerned with the question of whether the equilibrium is stable once attained.

[16] We thank Paul Weithman and an anonymous referee for bringing this point to our attention.

[17] We thank an anonymous reviewer for suggesting this example.

[18] For a balanced overview of cheap talk, see (Farrell and Rabin 1996). One might reasonably worry that noisy communication would not really be 'cheap' in a well-ordered society because informal social norms would impose costs on those who introduce sectarian, comprehensive reasons into public discourse. This, we believe, misunderstands the 'width' of the wide view of public reason. The wide view does not impose sanctions on persons for offering comprehensive reasons (so long as political reasons are not forthcoming), and so it imposes no sanctions on persons who would introduce noisy signals. There is good reason, then, to think that the introduction of such signals will be cheap, in lieu of some unknown norm to the contrary. We thank an anonymous reviewer for raising this concern.

[19] The risk dominant equilibrium in the game represented above in Note 7 is {Don't Comply, Don't Comply}. Technically speaking, players will move to the risk-dominant, sub-optimal equilibrium. This will only happen when the gains from the pay-off dominant equilibrium are not so high that assurance is unnecessary. We are claiming that 'cheap talk' leads to convergence on the risk-dominant equilibrium, only, as Aumann (1990) claims, that pre-play agreement or assurance based on 'cheap talk' is ineffectual. It cannot create the assurance we claim is necessary in the model public assurance game. Insofar as noise makes all public assurance 'cheap talk', we claim that this would publicly undermine assurance based on genuine signals that others will endorse and comply with the public conception. We thank an anonymous reviewer for helping us clarify this point.

[20] Technically speaking, the players will move to the risk dominant equilibrium.

[21] For instance, it must be admitted that cheap talk experiments tend to be run in small groups where little is at stake. Still, the experimental evidence is highly suggestive.

[22] For an influential account, see Bikhchandani et al. (1992).

[23] A correlated equilibrium, then, is a Nash Equilibrium for a 'super game' that includes at least three parties. In our example, these parties are drivers in perpendicular lanes and a traffic light. The traffic light or choreographer makes the first move. Then the other two players can do no better than obey the traffic light. In this case, the 'super game' is the two-step game where the choreographer moves and then the two drivers move.

[24] Conversation with Robert Sugden influenced our thinking on the importance salience (see also Sugden 1995).

[25] An anonymous reviewer wonders whether Rawls might be able to solve the assurance problem in an indirect way different from the one that we suggest. Rawls, in his reply to Habermas, writes that 'since there are far less doctrines than citizens, the latter may be grouped according to the doctrine they hold. More important than the simplifi-

cation allowed by this numerical fact is that citizens are members of various associations into which, in many cases, they are born, and from which they usually, though not always, acquire their comprehensive doctrines' (Rawls 1995: 15). Because there are only a small number of comprehensive doctrines, the reviewer suggests that this will simplify the assurance task, as Rawls suggests, and make assurance 'indirect' since it will be provided through groups of doctrines rather than by individuals directly. This is an intriguing suggestion and there is not space to address it completely here, but there are two reasons we think this version of 'indirect assurance' is not sufficient. Firstly, Following Gerald Gaus (2003: 180–6), we believe there are serious problems in individuating comprehensive 'doctrines' as a whole from sectarian reasons. If this fear is well-founded, the idea that members of a comprehensive doctrine can be organized in groups is misleading. Secondly, even if citizens can be grouped as the reviewer suggests, the fragility problem and the assurance problem reproduce at the level of the group. Members of the group will have to assure one another that they do, in fact, share doctrines for this 'indirect' solution to work. We believe this is unlikely but, again, a full reply would likely require another paper.

[26] The United States Supreme Court, for instance, provides reasons for its decisions, but they primarily affect future choreographing events rather than providing assurance to the public (the reasoning of the Supreme Court is very complex). We thank Micah Schwartzman for this point.

[27] A point that Rawls would, no doubt, admit. The point is the role that the choreographers play in the assurance mechanism.

[28] Members of the well-ordered society may have other mechanisms available, but Rawls only discusses the duty of civility, which is why we focus on it.

[29] A reviewer worries that debates about the interpretation of shared principles will not provide assurance unless persons are broadly committed to certain values and principles of constitutional interpretation. This assurance can be provided via choreography as well, specifically by honoring the adjudicative mechanisms that put constitutional interpretation into practices, such as the Supreme Court. We must proceed with care here, lest we fall back into the intuition that the only mode of assurance provision is through a commitment to shared reasoning rather than shared practices.

[30] For a sophisticated treatment of this conception of stability, see Kavka (1983).

[31] Ryan Muldoon has argued that diverse deliberative and non-deliberative settings can lead to robust public reasoning in the face of diversity (see Muldoon et al. 2012).

[32] The authors wish to thank many people who have commented on this paper or helped our thinking in conversation about this topic. Most notably: Jason Brennan, Jerry Gaus, Keith Hankins, Peter Leeson, Jonathan Quong, Micah Schwartzman, Robert Sugden, Kyle Swan and Paul Weithman.

REFERENCES

Anderson, L. and Holt, C. (1997), 'Information Cascades in the Laboratory', *American Economic Review*, 87: 847–62.

Aumann, R. (1990), 'Nash Equilibria are not Self-Enforcing', in J. J. Gabszewicz, J. F. Richard and L. Wolsey (eds) *Economic Decision Making: Games, Econometrics and Optimization*. Amsterdam: Elsevier.

—— and Hart, S. (2003), 'Long Cheap Talk', *Econometrica*, 71: 1619–60.

Bicchieri, C. (2002), 'Covenants without Swords: Group Identity, Norms and Communication in Social Dilemmas', *Rationality and Society*, 14: 192–228.
Bikhchandani, S., Hirshleifer, S. and Welch, I. (1992), 'A Theory of Fads, Fashion, Custom and Cultural Change in Informational Cascades', *Journal of Political Economy*, 100: 992–1026.
Farrell, J. and Rabin, M. (1996), 'Cheap Talk', *Journal of Economic Perspectives*, 10: 103–18.
Gaus, G. (2003), *Contemporary Theories of Liberalism: Public Reason as a Post-Enlightenment Project*. London: Sage.
Gintis, H. (2009), *The Bounds of Reason: Game Theory and the Unification of the Behavioral Sciences*. Princeton, NJ: Princeton University Press.
—— (2010), 'Social Norms as Choreography'. *Politics, Philosophy and Economics*, 9: 251–64.
Kavka, G. (1983), 'Rule by Fear', *Nous*, 17: 601–20.
Muldoon, R., Borgida, M. and Cuffaro, M. (2012), 'The Conditions of Tolerance', *Politics, Philosophy and Economics*, 11: 322–44.
Nussbaum, M. (2011a), 'Perfectionist Liberalism and Political Liberalism', *Philosophy and Public Affairs*, 39: 3–45.
—— (2011b), 'Rawls's Political Liberalism: A Reassessment', *Ratio Juris*, 24: 1–24.
Porter, D. and Smith, V. (1994), 'Stock Market Bubbles in the Laboratory', *Applied Mathematical Finance*, 1: 111–28.
Rawls, J. (1974), 'The Independence of Moral Theory', *Proceedings and Addresses of the American Philosophical Association*, 48: 5–22.
—— (1995), 'Political Liberalism: Reply to Habermas', *Journal of Philosophy*, 92: 132–80.
—— (1997), 'The Idea of Public Reason Revisited', *University of Chicago Law Review*, 64: 765–807.
—— (1999), *A Theory of Justice*, rev. edn. Cambridge, MA: Harvard University Press.
—— (2005), *Political Liberalism*, 2nd edn. New York: Columbia University Press.
Sally, D. (1995), 'Conversation and Cooperation in Social Dilemmas: A Meta-analysis of Experiments from 1958 to 1992', *Rationality and Society*, 7: 58–92.
Schelling, T. (1960) [1980], *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.
Skyrms, B. (2004), *The Stag Hunt and the Evolution of Social Structure*. New York: Cambridge University Press.
Smith, V. et al. (2008), 'Thar She Blows: Can Bubbles be Rekindled with Experienced Subjects?', *American Economic Review*, 93: 924–37.
Sugden, R. (1995), 'A Theory of Focal Points', *Economic Journal*, 105: 533–50.
Wilson, R. and Sell, J. (1997), ' "Liar Liar . . .": Cheap Talk and Reputation in Repeated Public Goods Settings', *Journal of Conflict Resolution*, 41: 695–717.
Weithman, P. (2010), *Why Political Liberalism? On John Rawls's Political Turn*. New York: Oxford University Press.