doi:10.1017/S0003055416000290

When Public Reason Fails Us: Convergence Discourse as Blood Oath

BRIAN KOGELMANN *University of Arizona*
STEPHEN G. W. STICH *University of Arizona and Yale Law School*

Public officials in John Rawls's well-ordered society face an assurance problem. They prefer to act in accordance with the political conception of justice, but only if they are assured that others will. On Paul Weithman's influential interpretation, Rawls attempts to solve this problem by claiming that public reason is an assurance mechanism. There are several problems with Rawls's solution: Public reason talk is too cheap to facilitate assurance, it is difficult to know when particular utterances express public reasons, and the requirements of public reason conflict with the fact of reasonable pluralism. We argue that convergence discourse—not public reason—solves the assurance problem by being a costly signal that indicates commitment to the political conception. This solution has none of Rawls's problems and has an interesting corollary: As diversity increases in society, so too does society's ability to solve the assurance problem. In short, the more diversity the better.

Those with a sense of justice aren't suckers. They wish to do right by their fellow citizens only if they are reasonably sure that their fellow citizens will do right by them. This presents an instability threat that some might find surprising: In a society filled with citizens who desire to act justly, everyone might act unjustly because they are unsure whether their fellow citizens will do the just thing. This is a basic assurance problem. John Rawls took seriously threats of instability to the well-ordered society and so was concerned about this basic assurance problem. Even after he showed that those in the well-ordered society would develop a sense of justice, Rawls still felt compelled to show that such a society would remain just. Part of doing this entailed showing how citizens in the well-ordered society assure each other that they will continue to act justly.

On one reading of Rawls's later thought, the assurance problem is solved by public reason. This solution might not be obvious, because public reason seems to have a normative purpose. The liberal principle of legitimacy, Rawls tells us, requires that we exercise political power in a manner justifiable to all. This creates a "moral, not a legal, duty"—the duty of civility—which requires that citizens "be able to explain to one another . . . how the principles and policies they advocate and vote for can be supported by the political values of public reason" (Rawls [1993] 2005, 217). But in addition to this moral function, many contemporary Rawls scholars—most notably, Paul Weithman and Stephen Macedo—believe that public reason also serves a more practical role. Specifically, these scholars believe that Rawls saw public reason as solving the assurance problem alluded to earlier.

In contrast, this article argues that public reason is incapable of solving the assurance problem. If political liberals wish to take seriously the threats of instability to a liberal order that so concerned Rawls, then they must find a new solution. This article offers just such a solution to the assurance problem in the form of a costly signaling model. Our main claim is that convergence discourse, the main theoretical competitor to public reasoning, is a costly signal capable of solving the political liberal's assurance problem. That is, convergence discourse succeeds where public reason fails. Our thesis comes with an interesting corollary: The ability of convergence discourse to solve the assurance problem is a positive function of how diverse the society is. In short, the more diversity the better. This corollary is in stark contrast to Rawls's understanding of diversity as a regrettable problem to be dealt with, not something to be celebrated (Rawls 1999, 12; Rawls 2001, 3–4). However, although we claim that convergence discourse is a more effective assurance mechanism than public reason, we do not claim that this is a conclusive argument for convergence discourse over public reason.

The structure of this article is as follows. In the next section we outline the two threats of instability faced by just societies and show how they are related to one another. Moreover, we show that Rawls's solution to these two instability problems is much more nuanced than those in the secondary literature have understood. From there we outline four criticisms of public reason as an assurance mechanism. We then present our own solution to the assurance problem and show that it does not generate those criticisms raised previously against public reason; this makes our convergence discourse model preferable to Rawls's public reason model. There is a concluding section.

TWO THREATS OF INSTABILITY

Just societies face two kinds of instabilities. Rawls illustrates both in this passage:

Brian Kogelmann is a Ph.D. student, Department of Philosophy, University of Arizona, Tucson, (bkogel89@gmail.com).

Stephen G. W. Stich is a Ph.D. student, Department of Philosophy, University of Arizona, Tucson, and a JD student, Yale Law School, New Haven, Connecticut (stich@email.arizona.edu).

The authors contributed equally to this work. They would like to thank those who attended their presentation at the Manchester Centre for Political Theory "Theories of Public Reason" panel in September 2015. Hun Chung provided helpful feedback on an earlier version of the article, and Jerry Gaus provided helpful discussion, as well as general guidance and mentorship. Finally, the authors would like to thank four anonymous reviewers and the editors of *American Political Science Review* for invaluable comments.

Figure 1. First Threat of Instability: The Prisoner's Dilemma.

	act on *P*	not act on *P*
act on *P*	3, 3	1, 4
not act on *P*	4, 1	2, 2

Figure 2. Second Threat of Instability: The Assurance Game.

	act on *P*	not act on *P*
act on *P*	3, 3	0, 2
not act on *P*	2, 0	1, 1

> Instability of the first kind is present when, if any person knows that the others will do their part, it will be worth his while not to do his: the consequences of one person's not doing his part if others do theirs may go unnoticed, or may have no ostensible effect, so that an alternative use of one's time and efforts is a personal gain.... Instability of the second kind is present when it is the case that if any one person knows or reasonably supposes that others will not do their part, it will be worth his while to be the first, or among the first, not to do his, or even dangerous for him not to be ([1963] 1999, 104).

Rawls's first threat of instability is best modeled as a Prisoner's Dilemma (Figure 1). Row's most preferred outcome is where Column acts on political conception of justice *P* and Row does not. The same applies for Column. Row and Column's second most preferred outcome is where both act on *P*, and the third most preferred outcome is where both do not act on *P*. The worst case outcome for Row is where Row acts on *P* and Column does not, and vice versa for Column. The only Nash equilibrium is (not act on *P*, not act on *P*).

Rawls's second threat of instability is best modeled as an assurance game (Figure 2). In this game, both Row and Column most prefer mutual adherence to *P*, a departure from the Prisoner's Dilemma. Row's worst case outcome is where Row adheres to *P*, but Column does not. The opposite is the case for Column. Therefore, if Row is not assured that Column will act on *P* and Column is not assured that Row will act on *P*, then both might not act on *P*. There are two Nash equilibria in this game, (act on *P*, act on *P*) and (not act on *P*, not act on *P*). In general, it is easier to solve assurance games than Prisoner's Dilemmas. With the assurance game, mutual adherence to *P* is already in equilibrium, and we just need to make sure that this equilibrium is realized and maintained. This is not the case with the Prisoner's Dilemma.

If we assume that societies are not polymorphic (not filled with multiple preference orderings), then both kinds of instability cannot exist at the same time. For simplicity we assume that we are working with non-polymorphic societies when it comes to preferences for acting on *P* and not acting on *P*. This means that these are two kinds of instability that obtain at different points in time: Society faces the Prisoner's Dilemma first and then the assurance problem.

When a society faces a Prisoner's Dilemma citizens most prefer to act unjustly when all other citizens act justly. How do we remedy this? By changing people's preferences. Rawls accomplishes this within his framework by arguing that citizens in the well-ordered society develop a sense of justice via a three-stage developmental process, which inculcates a desire to do what is just (Rawls [1963] 1999; Rawls 1971, ch. 8). When they desire to do what is just, citizens are no longer playing Prisoner's Dilemmas with one another. In Rawls's words, "For given these natural attitudes and the desire to do what is just, no one wishes to advance his interests unfairly to the disadvantage of others; this removes instability of the first kind" (1971, 497). When we change people's preferences the first kind of instability is mollified.

But even if everyone has a desire to act justly, we still face the second kind of instability: the assurance problem. As Rawls notes, "Even with a sense of justice men's compliance with a cooperative venture is predicated on the belief that others will do their part; citizens may be tempted to avoid making a contribution when they believe, or with reason suspect, that others are not making theirs" (1971, 336). Though citizens want to do what is just because of their sense of justice, they do not want to do what is just at any cost. Returning to Figure 2, if Column does not do what is just, then Row does not want to act justly and vice versa.

One might think that Rawls posits such deep levels of consensus and strong socialization in the well-ordered society that the assurance problem does not arise: In a society where everyone has a sense of justice and thus desires to act justly, why would citizens worry about the possibility of their fellow citizens acting unjustly and thus consider preempting such behavior by being the first to act unjustly? The key here is understanding the limits of Rawls's sense of justice: Although citizens with a sense of justice are social cooperators, they are not *unconditional* social cooperators. In Rawls's words, citizens "are ready to propose principles and standards as fair terms of cooperation and to abide by them willingly, *given the assurance that others will likewise do so*" (Rawls [1993] 2005, 49; emphasis added). This conditional nature of the desire to act justly suffices to ensure that there *is* an assurance problem *even in* the well-ordered society of deep consensus and socialization. All it takes to generate an assurance problem is a desire to not be a sucker, something even those with a sense of justice have. For there to be no such problem Rawls must say that those with a sense of justice are unconditional cooperators—that they desire to act justly *regardless* of what others do. It is obvious from the text that Rawls does not mean to make this claim.

How does Rawls then solve the assurance problem? He gives different answers over the course of his career—and, indeed, these different answers plausibly explain the changes from *A Theory of Justice* to *Political Liberalism*. Rawls's final answer to the assurance problem is that citizens, when deliberating in the public

sphere, adhere to the norms of public reason, which are norms governing public discourse (Hadfield and Macedo 2012; Weithman 2010, 327; Weithman 2015). Broadly, the norms of public reason prevent citizens from appealing to their comprehensive doctrines when engaged in public discourse. Rather, citizens may only appeal to the political conception of justice. When citizens adhere to the norms of public reason while engaged in public discourse, they signal commitment to the political conception of justice over their own comprehensive doctrine, assuring their fellow citizens that they will remain faithful to *P*. When citizens do not adhere to the norms of public reason they signal commitment to their comprehensive doctrine over the political conception, breaking down this assurance. Row thinks Column will remain loyal to *P* because Column adheres to the norms of public reason when engaged in public discourse and vice versa. Adhering to these norms signals fidelity to *P*.

There is a puzzle here that is not discussed in the literature. Implicitly we have assumed (along with the literature) that the assurance problem is a society-wide problem and that in our toy model Row and Column stand for any two random citizens. But according to Rawls, the norms of public reason only apply to a very small group of citizens engaged in a very small set of activities. He says that the idea of public reason applies to "the discourse of judges in their decisions, and especially judges of a supreme court; the discourse of government officials, especially chief executives and legislators; and finally, the discourse of candidates for public office and their campaign managers, especially in their public oratory, party platforms, and political statements" (Rawls [1997] 1999, 575). But if the norms of public reason only apply to very few members of society, how will they solve a society-wide assurance problem? If Row and Column are any two citizens who are not particularly politically involved, why would members of Congress debating legislation in accordance with the norms of public reason assure Row that Column will act on *P*, and vice versa?

It would not. But this is not a flaw with Rawls's solution to the assurance problem. Rather, the solution to the assurance problem is simply more complicated than those in the secondary literature let on. We can instead think of there being multiple assurance problems within different groups in society. These multiple assurance problems are differentiated by the strategies available to the participants; that is, how "act on *P*" is fleshed out. For high-ranking government officials, "act on *P*" means something like legislate according to the political conception of justice, decide court cases in accordance with the political conception of justice, and so on. If Row is a senator and Column a Supreme Court justice, how can Row be sure that Column will decide cases in accordance with the political conception? If Column does not do so, then Row prefers to legislate in accordance with her own comprehensive doctrine. And how does Column know that Row will legislate in accordance with the political conception? If Row does not then Column would like to decide major constitutional cases in accordance with his own comprehensive doctrine. We believe that Rawls means for the norms of public reason to solve *this* assurance problem, which we call the *public official assurance problem*. Because high-ranking public officials conduct public discourse in accordance with the norms of public reason they signal to other high-ranking government officials their fidelity to the political conception. On our interpretation of Rawls, the norms of public reason are meant to solve the public official assurance problem and *only* the public official assurance problem.

But, for average citizens, what does it mean to "act on *P*"? Because citizens do not have the capacity to legislate and decide cases, acting on *P* means that citizens simply obey laws whose content is in accordance with the political conception. Citizen Row wishes to follow just laws only if she is sure citizen Column will, and vice versa. How does Rawls solve this other assurance problem? He does not do so through the norms of public reason, because they do not apply. Important here is what Rawls says about the role of penal institutions in relation to stability. Rawls notes that "although men know that they share a common sense of justice and that each wants to adhere to the existing arrangements, they may nevertheless lack full confidence in one another." In remedying this problem, "the role of an authorized public interpretation of rules supported by collective sanctions is precisely to overcome this instability. By enforcing a public system of penalties government removes the grounds for thinking that others are not complying with the rules . . . the existence of effective penal machinery serves as men's security to one another" (Rawls 1971, 240). The *citizen assurance problem* is thus solved by penal institutions. How does citizen Row know citizen Column will follow the law? Because there are penal institutions that incentivize Column to do so. The same reasoning gives assurance to citizen Column.[1]

We think Rawls's solution to the citizen assurance problem is plausible.[2] As such, we want to focus on Rawls's solution to the public official assurance problem. A question: Since we think that penal institutions can solve the citizen assurance problem, why can they not solve the public official assurance problem? Because of the capacities that high-ranking government officials have—that is, what it means for them to "act on *P*"—there is no way for penal institutions

[1] Technically, there will also be a public official-citizen assurance game, where public officials must assure citizens that they will remain faithful to *P* by passing laws and deciding cases in accordance with *P*, and citizens must assure public officials that they will comply with these laws, if passed. We focus on the public official assurance problem in this article, and thus on how public officials assure one another. Although we believe that a model like ours could plausibly extend to the public official-citizen assurance game, the extension is not immediate and falls outside the scope of this article. We intend to address this third game in the future, but only claim here to provide a solution to the public official assurance game.

[2] Plausible but perhaps incomplete: Social norms will also play a large role in ensuring compliance and thus will do much to provide assurance. In the case of taxation, for instance, even when penal institutions exist, the number of citizens far exceeds the number of auditors, likely making the existence of penal institutions *alone* insufficient for providing assurance.

to ensure fidelity to the political conception of justice. For penal institutions to secure assurance among high-ranking government officials there must be laws forcing these public officials to reason in accordance with the political conception, to only pass laws in accordance with the political conception, and so on. Not only is it implausible for laws to actually do this, given the ambiguity in what the political conception of justice requires in terms of concrete legislative and judicial decisions, but *even if* laws could do this they would violate the basic liberty of free speech. So because penal institutions cannot solve the assurance problem among high-ranking government officials given the strategies available to these officials, the norms of public reason are required.

Failure to distinguish between the two different assurance games has prevented others in the literature from successfully solving the assurance problem. John Thrasher and Kevin Vallier, after arguing that public reason fails to solve Rawls's assurance problem (their criticisms, as well as criticisms given by others, are canvassed in the next section), offer their own unique solution to the assurance problem. On the Thrasher-Vallier model, assurance is provided in the well-ordered society so long as citizens follow "public choreographers," as well as see and believe their fellow citizens do so as well. According to Thrasher and Vallier, these "public choreographers are primarily bodies of norms, often legal, though sometimes informal or formal moral norms" (2015, 948). Of course, obeying *any* norms will not do—Row's obeying traffic laws will not assure Column that she will pay her taxes when the time comes, and hence act on *P* in the tax context. Rather, Row can provide this assurance only by paying her taxes or by obeying other norms related to tax compliance.

Thrasher and Vallier's assurance mechanism might replace or bolster the penal institution solution to the citizen assurance game. Yet it is hard to see how this solution can solve the public official assurance game given the nature of the strategies—what it means to "act on *P*"—available to the players. First, as we noted earlier, the relevant practices in this game—legislators legislating in accordance with the political conception of justice and judges judging in accordance with the political conception of justice—cannot plausibly be regulated by legal norms backed by penal institutions in a manner consistent with the basic liberty of free speech. But if there are no such norms regulating the relevant practices, then what does Judge Column obey that enables Senator Row to infer Judge Column's fidelity to *P*?

In response, perhaps there can be norms *not* backed by penal institutions (whether legal or informal) that regulate the relevant practices and thus do not violate the basic liberties. As an example, for judges, acting on *P* means judging in accordance with the political conception of justice. To do so, perhaps judges must follow interpretive rules that are not backed by formal sanctions: Say, they must adhere to originalism as an approach to legal interpretation. Part of acting on *P* thus involves following particular interpretive rules, and judges can signal fidelity to the political conception by following such norms.

But even here there are problems. Again, given the strategies available to the players in the public official assurance game, it will be difficult if not impossible to determine whether the relevant players are obeying the required rules. This is not so in the citizen assurance game: Citizen Row can tell rather simply if her fellow citizens are paying their taxes from existing data on the tax gap and, from there, infer whether her fellow citizens, on the whole, remain faithful to *P*. But can Senator Row read Judge Column's appellate-level decision and from there determine whether Judge Column obeyed the relevant interpretive norm? Given the prevalence of interpretive disputes even among those who espouse the same interpretive principles (such as originalism), this is incredibly unlikely, especially in cases decided by high-level appellate judges, which likely are the most significant for assurance purposes. As such, it is doubtful that players in the public official assurance game will be able to infer fidelity to *P* merely by witnessing how other players act on *P*, because whether a player has acted on *P* can be quite controversial.

THE FAILURES OF PUBLIC REASON

Does Rawls's public reason solution to the public official assurance problem succeed? Many do not think so. We agree, which is why we offer a new, discourse-based solution to the public official assurance problem later in the article. But before doing so, we outline several criticisms levied against public reason as an assurance mechanism in the existing literature and evaluate their cogency.

Too Cheap Talk

Gerald Gaus, followed by Thrasher and Vallier, criticizes public reason as an assurance mechanism in the following way (Gaus 2011, 317; Thrasher and Vallier 2015, 941–45). Suppose Row, by adhering to the norms of public reason in her discourse with Column, tries to signal to Column that she will act on *P* in order to induce Column to act on *P* as well, which is Column's best response to Row's acting on *P*. What should Column infer from this? Given his knowledge of Row's preferences, Column knows that it is in Row's interest *no matter what* to induce Column to act on *P*. If Column acts on *P*, the *worst* outcome Row can achieve is a payoff of 2. But if Column does not act on *P*, the *best* outcome Row can achieve is a payoff of 1. Given this, Column cannot infer from Row's adherence to the norms of public reason that Row will actually act on *P*. Whether or not Row actually plans to act on *P*, it is in Row's interest for Column to believe that Row will act on *P*, and thus it in Row's interest to adhere to public reason.

Implicit in this argument is the assumption that adhering to the norms of public reason is cheap talk: "We can understand Rawlsian displays of shared public reasoning as what economists call 'cheap talk'" (Gaus 2011: 317). Cheap talk is defined as communication that

does not affect the payoffs of a game (Farrell 1987, 35). Gaus thus assumes that adhering to public reason does not change Row's payoffs. If talk is not cheap, and if adhering to public reason changes Row's payoffs, then it might not be rational for Row to adhere to these norms if she plans to not act on *P*. If adhering to the norms of public reason is costly enough, it might only be rational to adhere to them if Row plans to act on *P*. If Column knows this, then Row's adherence to public reason *would* be sufficient to assure Column that she will act on P. We demonstrate this claim later in the article.

Technically, it is not necessary to assume that talk is cheap (in the way defined by economists) for this worry to be realized. Talk can still be costly, but not costly enough to change the structure of the game such that Row *only* has an incentive to adhere to the norms of public reason when she plans to act on *P*. The case of public reasoning as cheap talk is just a special, limit case of this worry. So this problem does not rely on the assumption that talk is cheap, but that talk is *too cheap*, by which we mean it is not costly enough to make it such that adhering to public reason is rational for Row and Column only if they plan to act on *P*.

Hence the fundamental question is as follows: Is adhering to the norms of public reason too cheap to render adherence to these norms rational only if one plans to act on *P*? We think that the answer to this question is yes, because we believe that adhering to the norms of public reason is cheap talk properly defined. Because we are working on the assurance problem, the assumption is that we are already in the well-ordered society and that we are concerned with how to *remain* in the well-ordered society—how do we continue to stay at the (act on *P*, act on *P*) equilibrium, rather than devolve to the (not act on *P*, not act on *P*) equilibrium? By definition, the well-ordered society is a society in which everyone knows and accepts, and knows that everyone else knows and accepts, the political conception of justice (Rawls 1971, 4–5). Since everyone knows the political conception of justice, giving reasons from this political conception in public discourse will not be costly. Legislators do not have to undergo the opportunity cost of learning the conception of the person as free and equal before they appeal to it in political debate, because they by hypothesis already know this—by assumption, it is part of the public political culture. As such, giving reasons in accordance with the political conception of justice as required by public reason is not costly. Because adhering to the norms of public reason is cheap talk, Row has reason to adhere to such norms when in discourse with Column whether or not she intends to act on *P*. Row's adherence will thus not assure Column at all. Call this the *too cheap talk problem*.

Common Knowledge

Gaus further criticizes public reason as an assurance mechanism by arguing that it requires *common knowledge*, which he thinks is implausible (2011, 317–18). Suppose first that the too cheap talk problem does not obtain. So when Row adheres to the norms of public reason she actually signals to Column that she will act on *P*, and vice versa. Suppose further that Row does this. Still, this is not sufficient to solve the assurance problem. Not only must Row signal to Column her fidelity to the political conception but Row must also know that Column has accurately received this signal, and Column must know that Row knows that Column received the signal, and so on and so forth ad infinitum. If Row signals to Column fidelity to the political conception, but is unsure whether Column has properly received her signal, then she might get spooked and think Column will not act on *P*, and thus she will also not act on *P* so to avoid her worst case outcome. And even if Row does know that Column accurately received the signal, Column might not know that Row knows this. So Column might get spooked that Row will not act on *P* and thus not act on *P* himself to avoid his worst case outcome. The problem iterates.

So we need common knowledge to adequately solve the assurance problem. Gaus argues that common knowledge in this setting is implausible: "Common knowledge is a very strong assumption . . . 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" (Gaus 2011, 318). Is Gaus's criticism of public reason as an assurance mechanism convincing? We think not, for reasons given earlier.

Previously we mentioned an important ambiguity in late Rawls's solution to the assurance problem: The assurance problem appears to be a society-wide problem, yet the norms of public reason are only meant to apply to high-ranking public officials in a very circumscribed setting. This led to our refinement of the assurance problem: There is the public official assurance problem, as well as the citizen assurance problem, and they are solved by different mechanisms. The norms of public reason are only meant to solve the public official assurance problem, not the citizen assurance problem.

Given that the norms of public reason apply only to the public official assurance game, the common knowledge objection fails. This is not because common knowledge is not needed in this new game—it is. Rather, the players in the public official assurance game *can* plausibly achieve common knowledge. Most people who study the realm of common knowledge hold that mutual witnessing of a public event among *n* persons is sufficient to generate common knowledge of that event among the *n* (Aumann [1976] 2000, 593; Milgrom 1981, 221). In the case of the public official assurance problem, adhering to public reason *is* a public event among the players. Appellate-level judges write opinions that political leaders must read when drafting and voting on statutes and that candidates for elected office must know in order to formulate viable policy platforms. Political leaders and candidates for elected office speak and debate. Often these communications are broadcast live and reported on by the media. Transcripts and recordings can be made available. In such cases Row *can* be sure that Column has received the message, and Column *can* be sure that Row knows this,

ad infinitum. When this is true there is no common knowledge problem.

So whether common knowledge is a problem for public reason as an assurance mechanism depends on to *whom* the norms of public reason apply. If the norms of public reason are meant to provide assurance for the citizen assurance problem, then Gaus is correct that it is unlikely that common knowledge could obtain, given the sheer size of the public. But if public reason applies only to the public official assurance problem, where the number of players is small and the game is played in a highly public manner, then it is plausible to think that common knowledge can obtain. As such, we do not think that common knowledge is a problem for public reason given our interpretation of the assurance problem. Because our new solution to the assurance problem is a proposed solution for the public official assurance problem only, we do not think common knowledge is a problem for our model either.

Noise

A third criticism of public reason as an assurance mechanism is the problem of noise (Thrasher and Vallier 2015, 941–45). Noise only becomes a problem when one adopts a specific interpretation of public reason. Under the more traditional view, public officials may only debate matters of basic justice and constitutional essentials with reasons taken from the political conception of justice. Call this the *exclusive* interpretation of public reason. Rawls later refines this view, holding that public officials can initially debate matters of basic justice and constitutional essentials with reasons taken from their comprehensive doctrines, so long as they back their respective positions up with public reasons eventually. Call this the *wide-scope* interpretation of public reason (Rawls [1997] 1999, §4).

The *noise problem* for wide-scope public reason is this. When Row is allowed to introduce reasons from her comprehensive doctrine in discourse with Column, then Column might be unsure if the reason Row gives is an actual public reason, even if Row intends it to be one. As Thrasher and Vallier note, "Once those other reasons are allowed [on the wide-scope view] . . . it will be difficult if not impossible to distinguish public reasons based on the public conception from those that are not so based" (2015, 942). As a consequence, *even if* there is no cheap talk and *even if* there is common knowledge, allowing wide-scope public reasons creates noise that can obfuscate genuine signals that Row and Column send to one another. Because Row cannot tell if Column's wide-scope reason is actually a public reason, she might be worried that Column will not act on *P*, forcing her to not act on *P* as well so as to avoid her worst case outcome.

Diversity

The final problem with public reason as an assurance mechanism is that it relies on a *consensus* view of political liberalism, rather than a *convergence* view of political liberalism (for an excellent overview of this distinction see Vallier 2011). This is a problem because consensus views of political liberalism may require an amount of agreement that is at odds with the fact of reasonable pluralism. At the very least, consensus views of political liberalism are less robust than convergence views, in that their models of the well-ordered society rely on both a greater number of assumptions and on more implausible assumptions than do convergence views.

Why does public reason as an assurance mechanism require a consensus view of political liberalism? To see why we first need to understand Rawls's three levels of publicity ([1993] 2005, 66–71). A society satisfies the first level of publicity when members of society know and accept the political conception of justice. Here we have mere convergence on a political conception of justice. A society satisfies the second level of publicity when members of society not only know and accept the political conception of justice but also adopt the "general beliefs about human nature and the way political and social institutions generally work, and indeed all such beliefs relevant to political justice" (Rawls [1993] 2005, 66). The third level of publicity is satisfied when citizens know and accept the full justification of the political conception of justice. For this third level of publicity to be satisfied, citizens must know and accept the method of political constructivism, how the original position decision procedure is set up, why the original position includes the features it does, and the like. When a society satisfies all three levels of publicity, agreement is quite deep, which is broadly what distinguishes consensus views from convergence views.

The three levels of publicity relate to the norms of public reason in that the sorts of facts about social science and human nature that citizens agree on when the second level of publicity is satisfied, as well as all those considerations agreed on when the third level of publicity is satisfied, give content to the norms of public reason (Rawls [1993] 2005, 68). As such, public reason as an assurance mechanism requires a consensus version of political liberalism, where agreement is quite deep. But one might object: Why do we need full publicity to get norms of public reason? Why can we not have norms of public reason with only mere convergence on a political conception of justice? The reason why the norms of public reason cannot be derived from the political conception itself is because our disagreements in the public sphere are often about how to best interpret our political conception of justice. For example, although we may all agree that the political conception requires freedom of speech broadly construed, it is unclear just what freedom of speech means when applied to specific cases. In such cases we cannot appeal to the political conception itself when debating what our political conception means. If we wish to avoid appealing to values from our comprehensive doctrines in debating such matters, then we need more fundamental values underlying our political conception to which we can appeal. Instead of appealing to our comprehensive doctrines we can appeal to those facts about human nature and social science we all

agree on as specified by the second level of publicity, as well as those considerations entailed by the third level of publicity. These considerations are public reasons. It is hoped that appealing to these sorts of facts and considerations will resolve the public debate in a way that offers assurance to fellow citizens.

The deep problem here is that those considerations entailed by the second and third levels of publicity can be quite controversial. Although it is up for debate just how controversial are the facts of social science and human nature entailed by the second level of publicity, certainly those considerations entailed by the third level of publicity are quite controversial. Indeed, this third level of publicity requires that citizens agree on the conception of the person as free and equal, what the two moral powers are, the method of political constructivism, and so on. It is implausible to think that such agreement can obtain in a society characterized by reasonable pluralism, because such matters are subject to the burdens of judgment. But without agreement here it is unclear where the content of our norms of public reason comes from—it is concepts like freedom and equality and the two moral powers that legislators and judges are supposed to appeal to when engaging in public reason. Using the norms of public reason as an assurance mechanism is thus inconsistent with reasonable pluralism: The conditions required for the assurance mechanism to succeed presuppose agreement that is at odds with the kind of diversity Rawls seeks to address. Call this the *diversity problem*.[3] The diversity problem is the final problem for public reason as an assurance mechanism.

COSTLY SIGNALS AS SOLUTIONS TO ASSURANCE PROBLEMS

Our solution to the public official assurance problem departs from the too cheap talk problem. According to this problem, Row's too cheap talk does not provide Column any assurance that Row will act on *P*. But what if talk is costly? As we now argue, costly talk solves the too cheap talk problem for the public official assurance problem, and convergence discourse is sufficiently costly to serve this function. To get a feel for the general solution, consider a simpler assurance game: the Stag Hunt.

The Stag Hunt (Figure 3) is an example of an assurance problem. Both players can either hunt stag or hunt hare. If they both hunt stag, then they work together; if a player hunts hare, then she hunts alone. Players successfully catch a stag only if they both hunt stag. If a player hunts hare, then she catches a hare regardless of what the other does. Given the amount of meat on each animal, each player's most preferred outcome is

Figure 3. The Classic Stag Hunt.

	Hunt Stag	Hunt Hare
Hunt Stag	5, 5	0, 3
Hunt Hare	3, 0	3, 3

(Stag, Stag), which is a strict Nash equilibrium. Given that each player hunts hare alone, and assuming that each player's hunting of hare has no effect on the other player's ability to catch a hare, Row is indifferent between (Hare, Hare) and (Hare, Stag). (Hare, Hare) is a nonstrict Nash equilibrium. The worst outcome for each player is to hunt stag while the other hunts hare. In this case she starves.

Suppose that talking is costless and that Row tells Column that she will hunt stag. What should Column infer? Nothing. Since talk is cheap, Row loses nothing by doing so. If Row plans to hunt stag then she should convince Column to hunt stag. If Row plans to hunt hare then she loses nothing by telling Column that she will hunt stag. So, once again, Column cannot infer from this talk what Row will do. This is the too cheap talk problem. Now suppose that, while telling Column that she will hunt stag, Row pulls out a knife and cuts her hand open. Why? Because cutting her hand open isn't cheap. Row is attempting to show Column that she really is committed to hunting stag. Column cannot infer a commitment on Row's part from too cheap talk. But why, if she plans to hunt hare anyway, would Row cause herself pain? She gains nothing from convincing Column to hunt stag if she plans to hunt hare. Thus, if she plans to hunt hare, cutting her hand open would be completely irrational. She would be imposing a strict loss on herself without any chance of a compensating benefit. But if Row plans to hunt stag then the behavior is rational. Row is willing to lower her utility to show Column that she is committed to hunting stag. The pain from the knife wound is a loss, but a loss that Row believes will be more than compensated by a successful stag hunt with Column. In other words, Column can infer that the knife cutting is an honest *costly signal* by Row.[4]

Of course, this might not actually motivate Row to hunt stag, because Row might not have sufficient assurance from Column. Consequently, when Row offers the knife to Column, there is good reason for Column to cut his hand open as well. This provides both players strong reason to believe that the other will hunt stag, because both players have given a costly signal that would make no sense were they planning to hunt hare, and this is common knowledge. Call this the *blood oath solution* to the assurance game. Notice that the blood oath solution works because the signal is costly and its

[3] Our diversity problem is similar to those who criticize public reason for being "incomplete," in the sense that it lacks sufficient content to resolve those debates that arise in the public sphere. See here Barry 1995, 144–45; Frohock 1997; Horton 2003, §4; Reidy 2000, 63–71; and Scanlon 2002, 163. Our diversity problem can be construed as saying that public reason is incomplete *precisely because* there is too much diversity to give sufficient content to the norms of public reason.

[4] Most scholars who have investigated costly signaling have applied it in the evolutionary biological context. See Grafen (1990) and Gintis, Smith, and Bowles (2001). For experimental examination of costly signaling see Aimone et al. (2013).

Figure 4. The cardinal public official assurance game.

	act on *P*	not act on *P*
act on *P*	10, 10	0, 3
not act on *P*	3, 0	2, 2

cost is common knowledge. As long as these conditions are met, it does not matter *why* the signal is costly. In particular, it is not necessary that the costs be meted out by a coercive authority via sanctions, as Rawls's solution to the citizen assurance game requires. All that matters is that there is *some* cost—how this cost is delivered is irrelevant to the blood oath solution's success.

The Stag Hunt is simpler than the public official assurance problem. In the Stag Hunt Row is indifferent between the Pareto-inferior Nash equilibrium (Hare, Hare) and the out-of-equilibrium solution that she prefers (Hare, Stag). Consequently, it is *necessarily* irrational for her to take the blood oath in order to convince Column to play the out-of-equilibrium solution. It imposes a loss on her for which there is no compensating benefit. So there is no reason for her to take the blood oath if she plans to violate it. Thus there is no reason for Column to be suspicious of *any* costly signal Row sends, no matter the cost. But this is not so in the public official assurance problem. This is because Row prefers the out-of-equilibrium solution (not act on *P*, act on *P*) to the Pareto-inferior Nash equilibrium (not act on *P*, not act on *P*). So there *could* be a reason for Row to take a blood oath to act on *P* even if she plans on violating the oath. She could plan to not act on *P* and hope to convince Column, by taking the blood oath, to act on *P*. Knowing all this, Column should still be suspicious. Row could simply be engaging in too cheap talk even while engaging in costly signaling.

Solving this problem requires resorting to cardinal utilities. Figure 4 is a cardinal public official assurance game with the same ordinal structure—hence, the same pure-strategy Nash equilibria—as the public official assurance game in Figure 2. But the utility each player gains from the (act on *P*, act on *P*) equilibrium is far superior to any other option. This makes a difference.

Suppose the blood oath to act on *P* costs Row a utility of 0.5. This behavior could be rational no matter how Row plans to act. If she plans to act on *P* and successfully convinces Column to act on *P*, then she gains a utility of 9.5 compared to the outcome that otherwise would have obtained—where she acts on *P* but Column does not (because her payoff in this case is zero). But if Row plans to not act on P and successfully convinces Column to act on *P*, then she has gained 0.5 utility compared to the outcome that otherwise would have obtained—where both Row and Column do not act on *P* (because her payoff in this case is two). No matter what, Row gains utility. But now suppose that cutting her hand open costs Row utility *s*, where $1 < s < 8$. In this case, cutting her hand open is rational only if Row plans to act on *P*. For suppose Row plans to not act on *P*. Then Row has caused herself more pain than she gains from Column's switch from not acting on *P* to acting on *P*. Given that this would be irrational, Column should conclude that Row's portion of the blood oath is sincere when the blood oath costs Row utility *s*.

The ordinal structure of the public official assurance game guarantees that Row will be able to send such a signal, assuming that Row believes with sufficient confidence that her signal will induce a switch on Column's part from not acting on *P* to acting on *P*. To communicate that she intends to act on *P*, it suffices for Row to sacrifice more than she would gain by (1) fooling Column to act on *P* and (2) not acting on *P* herself. There will *always* be some such signal so long as the following inequality holds, letting u_{Row} indicate utility to Row: u_{Row}(act on *P*, act on *P*) – u_{Row}(act on *P*, not act on *P*) > u_{Row}(not act on *P*, act on *P*) – u_{Row}(not act on *P*, not act on *P*). This inequality says that Row gains more utility from inducing Column to switch from not acting on *P* to acting on *P*, *given that Row will act on P*, than Row gains from inducing Column to switch from not acting on *P* to acting on *P*, *given that Row will not act on P*. The satisfaction of the inequality suffices for the existence of a possible effective signal because the value of an effective signal is just some positive number *s* such that left-hand side > *s* > right-hand side. Thus, if left-hand side > right-hand side then, trivially, there exists some *s*. That this inequality will hold is guaranteed by the ordinal structure of the game because u_{Row}(act on *P*, act on *P*) is greater than u_{Row}(not act on *P*, act on *P*), and u_{Row}(not act on *P*, not act on *P*) is greater than u_{Row}(act on *P*, not act on *P*): The first term on the left-hand side is greater than the first term on the right-hand side, and the second term on the left-hand side is less than the second term on the right-hand side. Therefore, there will always be some amount that Row can spend that would signal her sincerity, assuming she believes that her signal will be effective. The same holds for Column.

CONVERGENCE DISCOURSE IS LIKE CUTTING YOUR HAND OPEN

So far we have proceeded at a high level of abstraction. We have shown that there will be some available costly signal that Row can send to assure Column, but we have not shown how Row can send such a signal. This section addresses this issue. The guiding idea is that *convergence discourse is a costly signal*.

First, it is important to be clear on what convergence discourse requires and how it likely proceeds. Public officials advance ideas in the public political forum. On consensus political liberalism models, public officials must justify these ideas to others only by appealing to public reasons taken from the political conception of justice. On convergence political liberalism models, conditions are much more lax: The set of considerations public officials may permissibly appeal to is much larger

and includes reasons from public officials' comprehensive doctrines. That said, convergence models of public discourse do not say that public officials may appeal to *any* considerations *whatsoever*. Vallier, for instance, imposes an *intelligibility restraint*, which requires that Row only appeal to reasons in her discourse with Column that are epistemically justified to Row according to her own evaluative standards (Vallier 2014, 183–85).[5] But, as Vallier notes, because many citizens' arguments will survive this test, the "restraint will apply to relatively few reasons" (182). So although convergence discourse places some restraints on the considerations to which public officials may appeal, these restraints are quite permissive when compared to public reason. If Row is a Christian, for instance, she may appeal to Christian-based reasons from her comprehensive doctrine. On standard accounts of public reasoning she may not do this. And under the least restrictive wide-scope interpretation of public reasoning discussed earlier, Row may appeal to her Christian-based reasons only if she can back up her position with public reasons in due time. With convergence discourse Row may appeal to Christian-based considerations full stop.

Of course, appealing *only* to reasons from one's comprehensive doctrine as convergence discourse permits will likely result in little success for public officials. If Row is a Christian and Column a Hindu, and Row wishes to convince Column to endorse policy p, then merely giving Column Christian-based reasons to endorse p will do very little to convince him. Moreover, such discourse is at odds with having a sense of justice. According to Rawls, citizens in the well-ordered society "express a willingness, if not the desire, to act in relation to others *on terms that they also can publicly endorse*" (Rawls [1993] 2005, 19; emphasis added). But Column, a Hindu, cannot publicly endorse Row's Christian-based reasons. For public officials engaged in convergence discourse to (1) successfully convince others to support their respective positions and to (2) also engage with their fellow public officials on terms their fellow public officials can endorse, they will have to engage in what Rawls calls *reasoning by conjecture*. When reasoning by conjecture "we argue from what we believe, or conjecture, are other people's basic doctrines, religious or secular, and try to show them that, despite what they might think, they can still endorse a reasonable political conception" (Rawls [1997] 1999, 594).

In convincing Column that p is a good policy, Row will thus give Column Hindu-based reasons. To succeed at this effort Row must come to learn Column's comprehensive doctrine to a significant degree. If she does not know Column's comprehensive doctrine well, then she will not be able to show Column that Column's comprehensive doctrine entails p, nor will she be able to engage with Column on terms Column can endorse. Note that this style of reasoning is not a *requirement* of convergence discourse, but rather is a *likely feature* of it given (1) what it takes to convince diverse persons with no shared reasons to come to agree with one's position and (2) what convergence discourse likely looks like when engaged in by persons with a sense of justice as defined by Rawls.

In short, successful participation in convergence discourse in the well-ordered society likely requires knowing a significant amount about a wide range of comprehensive doctrines that are not one's own, which can be costly in terms of the opportunity cost spent learning the relevant doctrines. Row, whom Column does not expect to know much about Hinduism, has given a sophisticated Hindu-based argument. Because Row is not a Hindu, it must have taken her a great deal of effort to learn the doctrine sufficiently well to give this argument. That Row would be willing to incur such a cost indicates that she is serious about achieving the (act on P, act on P) outcome. Row probably would not have incurred such a cost if she merely wanted to secure the (not act on P, act on P) outcome, just as one would not likely cut one's hand open in the Stag Hunt if one planned on hunting hare. As such, if Row makes a compelling Hindu-based argument in favor of p, then Column should be assured that Row desires this outcome. As in the blood oath, costly signals work best if they go both ways. Row still needs assurance that Column will act on P. Column can so assure Row by giving Christian-based arguments for his preferred policy.

We just said that, because Row is not a Hindu, it must have taken her a great deal of effort to learn the argument based on Hinduism to convince Column. But is this effort sufficiently costly to overcome the too cheap talk worry? After all, although we argued earlier that it is a necessary feature of the assurance game that there will be *some* sufficiently costly signal Row can send, we have not shown that convergence discourse is costly enough. We now argue that the likelihood that this effort is sufficiently costly to overcome the too cheap talk worry is a positive function of society's diversity of comprehensive doctrines. More bluntly, *the more diverse the society, the more effective the blood oath solution*. To see this, consider two cases.

Small Diversity. There are two comprehensive doctrines: Christianity and Hinduism. Row therefore can produce sufficient reasons for all members of society to endorse the (act on P, act on P) outcome by knowing arguments from these two comprehensive doctrines.

Large Diversity. There are many comprehensive doctrines, including Buddhism, Christianity, Hinduism, Islam, Jainism, Judaism, Kantian Liberalism, Libertarianism, Millian Liberalism, Secular Humanism, Taoism, and Zoroastrianism. Row therefore must learn arguments from most or all of these comprehensive doctrines to convince the others to endorse the (act on P, act on P) outcome.

[5] Vallier (2014) places slightly more stringent restraints on *proposals* that public officials may advance in the public sphere, although even here these restraints are still quite lax. For simplicity we ignore this and focus on restraints on *reasons*, or restraints on considerations that public officials may appeal to when trying to convince other public official that their position is the correct position.

The signal from Row's convergence discourse is significantly more costly in the large diversity case than in the small diversity case. She must expend many more resources to learn arguments sufficient to convince others that their comprehensive doctrine leads them to endorse the (act on P, act on P) outcome simply because there are many more such arguments to learn. This example shows that, the more comprehensive doctrines there are to learn, the greater the chance that Row's costly signal can overcome the too cheap talk problem. Thus, although Rawls views diversity as creating a stability problem that must be solved, we conclude that *diversity is an integral part of the solution to this very same problem*. It is an integral part of the solution to the too cheap talk problem. We thus join a growing literature in political philosophy and the philosophy of science that seeks to show the social benefits that diversity brings to the table (D'Agostino 2009; 2010; Muldoon 2013; Muldoon and Weisberg 2011; Page 2008).

So far we have assumed for the sake of simplicity that the public official assurance game is a one-shot interaction. However, often this is not the case. Public officials interact over long periods of time: The game is *iterated*. We have also argued that convergence discourse is costly because it requires public officials to learn about many comprehensive doctrines. Yet these costs are mostly incurred up front. Once Column has learned enough Christian doctrine to give one Christian-based argument, giving Christian-based arguments in the future will be quite cheap—perhaps insufficiently costly to avoid the too cheap talk problem. So, large diversity is necessary but not sufficient for convergence discourse to be sufficiently costly when playing an iterated public official assurance game. Now consider the following condition:

> *Dynamic Large Diversity*. There are many comprehensive doctrines in society at time t_1. There are many comprehensive doctrines in society at time t_2. The t_1 doctrines are significantly different from the t_2 doctrines due to the evolution and reinterpretation of existing comprehensive doctrines, as well as the introduction of new doctrines. At t_1, Row must learn arguments from the t_1 doctrines to convince the others at t_1 to endorse the (act on P, act on P) outcome. Likewise with the t_2 doctrines at t_2.

Public officials in conditions of dynamic large diversity will maintain the costliness of their signals through iterated interactions because they must pay attention to cutting-edge doctrinal developments to (1) successfully convince others to support their respective positions and to (2) also engage with their fellow public officials on terms that their fellow public officials can endorse. This ensures the costliness of convergence discourse over time. Moreover, public officials will be able to accomplish this by citing recently published sources. For example, Column could cite the most recent papal encyclicals for Row, which are produced fairly regularly. So the blood oath solution solves the iterated public official assurance problem in conditions of dynamic large diversity.

Is it plausible that liberal societies such as Rawls's well-ordered society will contain dynamic large diversity? Yes. As Ryan Muldoon has emphasized, dynamic large diversity is the natural result of liberal institutions because liberal institutions enable persons to experiment with different ways of living. According to Muldoon, "[t]hese ways of living may embody different perspectives on how we should live together in a society" (2015, 193). Because "the process of experiments in living is continuous," so too is the evolution of perspectives—and hence comprehensive doctrines—present in the society (180). Just as Rawls believed that the very liberal institutions constituting the well-ordered society would lead to a diversity of comprehensive doctrines in the first place (Rawls [1993] 2005, 36–37), Muldoon rightly shows that these same institutions lead to a *shifting* diversity of comprehensive doctrines. We should thus expect Rawls's well-ordered society to contain dynamic large diversity. This allows the blood oath solution to solve the iterated public official assurance game.[6]

Still, theorists who show the benefits of diversity also argue that diversity has its limits. We agree. Return to Figure 4 and suppose that learning one comprehensive doctrine costs players utility 1. In this game it is rational for players to learn at most eight comprehensive doctrines: Were they to learn nine they would get a maximum payoff of 1, but not acting on P without costly signaling guarantees a payoff of at least 2. Now suppose there are 16 comprehensive doctrines (*CD*s) in the society, that Row knows CD_1 through CD_8, and that Column knows CD_9 through CD_{16}. Row cannot send an effective costly signal because Column, not knowing the doctrines through which it is expressed, cannot verify whether Row expended significant resources in learning the doctrines. Here, diversity is too great.

Although situations like this are possible, we do not think they are plausible in the well-ordered society. Thus far we have modeled the public official assurance game between two players, but in reality the public official assurance problem has n players, where n is (probably) in the hundreds and players must perceive some number $q < n$ of officials signaling to be sufficiently assured: Row does not need assurance from Column so long as she perceives q players costly signaling. Now

[6] There are other plausible mechanisms to solve the iterated public official assurance problem, mechanisms that could work *even without* dynamic large diversity. One such mechanism is costly signals that persist in costliness. Following Gambetta (2009), consider facial tattoos used for gang induction. These are not signals whose costs are strictly upfront, but rather persist over time: They signal commitment to the gang *for life*, because the tattoo removes traditional employment options from the feasible set *for life*. That is, the signal continues being costly. It is possible that convergence discourse is a costly signal of this kind, although the nature of the persisting cost is different from the nature of the upfront cost. For example, alhough the opportunity cost of learning Muslim doctrine is purely upfront for Rick Santorum, being the sort of person who makes arguments appealing to the Koran imposes a cost that may persist over time, given the nature of Santorum's constituency.

add a third player, Matrix, who knows CD_5 through CD_{12}, to the game, and suppose $q = 2$. If each player publicly discusses at least six comprehensive doctrines then q is necessarily satisfied for each player. Row and Column each perceive two costly signals (themselves and Matrix), and Matrix perceives three costly signals.

Matrix is a *connection player* between Row and Column. As the last example illustrates, connection players can facilitate assurance even when most players know no common comprehensive doctrines. Thus, connection players increase the amount of diversity a society can afford to have. Various properties of the well-ordered society suggest that the number and effectiveness of connection players will be high. First, the basic liberties governing the well-ordered society, such as freedom of speech and movement, make likely a large number of connection players because they facilitate significant social interchange (Rawls 1993 [2005], 228).

Second, connection players are effective in proportion to the number of comprehensive doctrines they can rationally learn: If Matrix could learn only two comprehensive doctrines then she could not send effective costly signals to both Row and Column. A player's maximum rational number of known comprehensive doctrines depends on her cardinal preferences over the outcomes. In particular, if Row *strongly* prefers (act on P, act on P) to the other outcomes, then she can rationally learn many comprehensive doctrines—she gains *a lot* from inducing Column to act on P and so can rationally send a very costly signal. To see this, compare Figure 4 with Figure 2, and let Figure 2 now model cardinal utilities. Assuming that learning a comprehensive doctrine costs utility 1, players can rationally learn at most eight doctrines if they are playing the Figure 4 game and at most two if they are playing the Figure 2 game.

It is impossible to specify a priori the players' cardinal utilities, which will depend on features particular to individual public officials. However, given that the relative cardinal utilities of the outcomes make a big difference to the well-ordered society's ability to solve the public official assurance problem, we should expect the well-ordered society to inculcate favorable preferences in its public officials as part of the general developmental process outlined by Rawls ([1963] 1999; 1971: ch. 8). For example, Matrix's commitment to achieving justice plausibly leads her to strongly prefer (act on P, act on P) to the other outcomes, thereby increasing her effectiveness as a connection player. Thus, even though too much diversity is a theoretical possibility, we doubt it would obtain in the well-ordered society.

OBJECTIONS

Public Goods

It might be thought that the preceding paragraphs raise a public goods problem. We just argued that public officials need not perceive the costly signals of *all* other public officials, but just some q, to be sufficiently assured. This claim suggests that assurance can be achieved even if not every player signals. Because not every player must signal, perhaps it is rational for each player not to signal and then either act on P or not act on P depending on how many signals she observes. The result: Nobody signals and nobody acts on P.

On closer inspection, the public goods problem does not arise. Nonsignaling is an instance of free-riding. Nonsignalers free-ride off the activities of signalers. Now, recall the first instability threat that just societies face: the Prisoner's Dilemma. A public goods problem has the same structure as an n-person Prisoner's Dilemma: Row's best outcome is to free-ride off of Column's cooperation (signaling) by defecting (not signaling). Recall how Rawls solves this problem: He argues that citizens in the well-ordered society will have a sense of justice and that citizens with a sense of justice do not play Prisoner's Dilemmas with one another. No such person most prefers to "advance his interests unfairly to the disadvantage of others" (Rawls 1971, 497). Citizens prefer to act fairly in these matters. But free-riding off of others' costly signaling is a paradigm case of advancing one's interests unfairly to the disadvantage of others; namely, those who send costly signals. Therefore, persons in the well-ordered society will not have preferences that give rise to a public goods problem.

Not Costly Enough

Next, it might be objected that facts about everyday politics impede signals from being sufficiently costly. For example, to effectively reason by conjecture, political officials might hire experts in particular comprehensive doctrines to make the relevant arguments before political bodies: Row might simply hire a Hindu to testify before Congress and not learn about Hinduism herself. Public officials could thus ensure that the audience is exposed to these arguments at minimal or no cost. Nothing in the minimal requirements of convergence discourse precludes this behavior.

Yet this problem can be solved through appropriate rules governing deliberative assemblies. Consider *Robert's Rules of Order*, which are rules of order governing discourse among deliberative assemblies. Although these rules of order (or some version thereof) are applied widely in contemporary liberal democracies, they are not the only possible rules that can be used. If there is a worry that political officials can avoid costly signaling by having experts come testify before Congress, then certain rules of order can mitigate these fears. For instance, there can be a rule governing legislative assemblies holding that experts may only testify on matters of science or social science in committees, but not on normative matters. Though this dividing line is no doubt blurry and will be subject to hard cases, such a rule of order would prevent Row from bringing in a Hindu to come give Column Hindu-based reasons in favor of policy p. Because the rules of order prevent Row from doing this, to muster a convincing argument in defense of p Row must undergo the opportunity

cost and learn some information about Hinduism. This preserves the costliness of Row's signal.

But even if Row's signal is sufficiently costly in that she underwent the opportunity cost, will Column perceive it as such? Even if Row gives enough arguments from other comprehensive doctrines, Column might not be able to tell whether Row actually expended significant resources to learn comprehensive doctrines that are not Column's. For all Column knows, Row is just using buzzwords from these comprehensive doctrines and has not actually expended any significant resources in learning them. Or perhaps Row could have hired someone to come up with these arguments for her behind the scenes and has just memorized them. Although this effort would cost something, it might not cost a lot. So even if Row in fact did expend these resources, this expenditure is not an effective costly signal.

We have two responses to this sort of objection. First, to the buzzwords point: As a participant in the convergence discourse, Column himself has learned these doctrines to give arguments from them. So Column actually *can* evaluate Row's arguments from these different comprehensive doctrines. He can tell the difference between buzzwords and good arguments and can thus verify whether Row took the time to learn the arguments and thus to send a genuine costly signal.

Second, hiring others to write her arguments for her could only get Row so far. As we argued earlier, the individuals involved in the public official assurance problem are, inter alia, appellate-level judges, high-level political officials, and candidates for elected office. Granted, sometimes these persons need only read speeches, for which they require no knowledge of others' comprehensive doctrines. But they also engage in live, public debate-esque activities in which they will be asked questions about others' comprehensive doctrines. They will need to be able to answer these questions. To be able to do so, they will need to be able to think on their feet using premises from others' comprehensive doctrines. And to do this they will need knowledge of these doctrines. Thus, if Row is a public official, she will need knowledge of many comprehensive doctrines to succeed in these activities. For these reasons, Column can infer that Row's signal is costly.

Too Demanding

Our responses to the last objection raises another worry—that what we are asking of public officials is too demanding: It asks too much of Row to learn all the relevant comprehensive doctrines in order to successfully argue for p. Our response: Although this effort is quite demanding, it is permissibly so given the particular assurance game to which it applies. If our costly signaling model was meant to apply to the citizen assurance game, then we agree it would be problematically demanding—asking *all* citizens qua citizens to learn many comprehensive doctrines is implausible. But our costly signaling model only applies to the public official assurance game, where citizens freely take up the respective duties and responsibilities of office. Because these duties and responsibilities are voluntarily taken up and are thus supererogatory, we do not think we ask too much. Here is an analogy: Being an investment banker is a very demanding career, but many people do not think it is problematically so. Why? Because people are not forced to be investment bankers. Similarly, structuring political institutions so that public officials are forced to costly signal is not too demanding, because citizens can live good and fulfilling lives without ever entering the public sphere. When citizens *do* choose to join the public official assurance game, they do so knowing the demands of office that, we agree, can be quite demanding on our model.

Noise

Recall the four problems with Rawlsian public reason: too cheap talk, common knowledge, noise, and diversity. We already showed that the costly signaling model does not have problems with too cheap talk and diversity. First, the model was designed to make talk sufficiently costly, and indeed, that is why the model works. Second, diversity is not a problem for our model. On our model, the more diversity the better. Moreover, because our costly signaling model is only meant to solve the public official assurance problem, it, like Rawls's public reason model, also does not run into the common knowledge objection. All that remains is noise.

Thrasher and Vallier argue that "assurance provided by direct public reason can be undermined by noise" (2015, 941). In direct public reason, "citizens' direct deliberation with one another provides assurance" (941). By "direct public reason" Thrasher and Vallier thus include convergence discourse, because convergence discourse is a form of direct deliberation. So it might be thought that our costly signaling model is subject to the noise problem, because Thrasher and Vallier seem to be claiming that even convergence discourse suffers from noise. However, Thrasher and Vallier are of two minds about the ability of noise to undermine *all* "direct public reason," because they claim that "[t]he nice feature of the exclusive view [of public reason] is that it prevents the 'noise' associated with comprehensive doctrines from undermining mutual assurance" (940). Since, by their definitions, exclusive public reason is an instance of "direct public reason," this last claim implies that noise does *not* actually undermine all direct public reason, contra what they claim in the first half of this paragraph.

We agree with the last claim. For noise to be a problem for any type of direct public reason two facts must obtain: (1) The discourse is noisy, and (2) deliberators cannot cure noise. Thrasher and Vallier show that both (1) and (2) apply to the wide-scope interpretation of public reason. But (1) does not apply to exclusive public reason, in which persons may *only* introduce shared reasons taken from the political conception of justice. This, presumably, is why Thrasher and Vallier claim that noise does not undermine exclusive public reason.

Figure 5 Noisy cardinal public official assurance game.

	act on *P*	not act on *P*
act on *P*	10, 10	0, 9
not act on *P*	9, 0	2, 2

However, (2) does not apply to convergence discourse. As Thrasher and Vallier point out, the underlying problem with noise is that it increases the uncertainty that given utterances are the relevant assurance-giving signal. For the wide-scope interpretation of public reason, the relevant assurance-giving signal is discourse based on public reasons, which Thrasher and Vallier argue cannot be easily distinguished from nonpublic reasons; hence, deliberators using wide-scope public reason cannot cure noise. For convergence discourse, the relevant assurance-giving signal is a costly signal. So, if there is a noise problem for convergence discourse, the problem is how to distinguish sufficiently costly signals from insufficiently costly signals (i.e., too cheap talk).

Define the "too cheap talk baseline" b as the maximum amount it would be rational for Row to signal, given that Row plans to not act on P. In Figure 4, for example, $b = 1$; in the Stag Hunt, $b = 0$. Let s be a signal's cost, and let s_{Max} be Row's maximum rational costly signal. Plausibly, the greater the difference between s and b, the easier it is to distinguish assurance-giving signals (i.e., costly signals) from non-assurance-giving signals (i.e., too cheap talk), and hence the less noisy the discourse: In the Stag Hunt, discourse is less noisy if players cut off a finger or two than if they barely draw blood. So convergence discourse can cure noise if players signal significantly more than the too cheap talk baseline.

Whether it is rational for players to do so depends on their cardinal preferences over the outcomes. For example, in Figure 4, $b = 1$ and $s_{Max} = 8$. If the players send fairly high signals—say, $s = 5$—then they could easily identify them as costly and avoid the noise problem. However, in Figure 5, $b = 7$ and $s_{Max} = 8$. Given that no rational signal can be much costlier than the too cheap talk baseline, there is little Row can do to prevent noise.

The elimination of noise thus depends on $s_{Max} - b$ being large, which is a function of the players' cardinal preferences. Note that $s_{Max} = u_{Row}$(act on P, act on P) $- u_{Row}$(not act on P, not act on P), and $b = u_{Row}$(not act on P, act on P) $- u_{Row}$(not act on P, not act on P). Simplifying the difference, $s_{Max} - b = u_{Row}$(act on P, act on P) $- u_{Row}$(not act on P, act on P). Thus, there is a large difference between s_{Max} and b if and only if there is a large difference between u_{Row}(act on P, act on P) and u_{Row}(not act on P, act on P). We already argued earlier that Row's sense of justice will render u_{Row}(act on P, act on P) high. This sense of justice will also render u_{Row}(not act on P, act on P) low, because in this case she takes advantage of Column's acting justly, which those with a sense of justice are not inclined to do. Thus, in the well-ordered society, the difference between s_{Max} and b is high, implying that the blood oath solution can avoid the noise problem.

A CONCLUDING SECTION

Individuals in Rawls's well-ordered society must assure one another that they will act in accordance with the political conception of justice. Rawls attempts to solve the citizen assurance problem with penal institutions. He succeeds. Rawls attempts to solve the public official assurance problem with public reason. He fails. Public reason as an assurance mechanism faces the problems of too cheap talk, noise, and diversity. We have proposed a different solution to the public official assurance problem: Convergence discourse is a costly signal akin to a blood oath. We have succeeded. Our solution has none of the problems that Rawls's does and is plausible more generally. The lessons to be drawn from our model are twofold. First, convergence discourse, unlike public reasoning, solves the public official assurance problem. Second, the existence of significant diversity is necessary for convergence discourse to succeed. Diversity solves the problem that diversity creates.

REFERENCES

Aimone, Jason, Laurence Iannaccone, Mike Makowsky, and Jared Rubin. 2013. "Endogenous Group Formation via Unproductive Costs." *Review of Economic Studies* 80: 1215–36.

Aumann, Robert. [1976]2000. "Agreeing to Disagree." In *Collected Papers, Vol. 1*. Cambridge, MA: MIT Press, 593–96.

Barry, Brian. 1995 *Justice as Impartiality*. Oxford: Clarendon Press.

D'Agostino, Fred. 2009. "From the Organization to the Division of Cognitive Labor." *Politics, Philosophy and Economics* 8: 101–29.

D'Agostino, Fred. 2010. *Naturalizing Epistemology: Thomas Kuhn and the "Essential Tension."* New York: Palgrave Macmillan.

Farrell, Joseph. 1987. "Cheap Talk, Coordination, and Entry." *RAND Journal of Economics* 18: 34–39.

Frohock, Fred M. 1997. "The Boundaries of Public Reason." *American Political Science Review* 91: 833–34.

Gambetta, Diego. 2009. *Codes of the Underworld: How Criminals Communicate*. Princeton: Princeton University Press.

Gaus, Gerald. 2011. "A Tale of Two Sets: Public Reason in Equilibrium." *Public Affairs Quarterly* 25: 305–25.

Gintis, Herbert, Eric Alden Smith, and Samuel Bowles. 2001. "Costly Signaling and Cooperation." *Journal of Theoretical Biology* 213: 103–19.

Grafen, Alan. 1990. "Biological Signals as Handicaps." *Journal of Theoretical Biology* 144: 517–46.

Hadfield, Gillian K., and Stephen Macedo. 2012. "Rational Reasonableness: Toward a Positive Theory of Public Reason." *Law and Ethics in Human Rights* 6: 7–46.

Horton, John. 2003. "Rawls, Public Reason, and the Limits of Liberal Justification." *Contemporary Political Theory* 2: 5–23.

Milgrom, Paul. 1981. "An Axiomatic Characterization of Common Knowledge." *Econometrica* 49: 219–22.

Muldoon, Ryan. 2013. "Diversity and the Division of Cognitive Labor." *Philosophy Compass* 8: 117–25.

Muldoon, Ryan. 2015. "Expanding the Justificatory Framework of Mill's Experiments in Living." *Utilitas* 27: 179–94.

Muldoon, Ryan, and Michael Weisberg. 2011. "Robustness and Idealization in Models of Cognitive Labor." *Synthese* 183: 161–74.

Page, Scott. 2008. *The Difference*. Princeton: Princeton University Press.

Rawls, John. [1963]1999. "The Sense of Justice." In *Collected Papers*. Cambridge, MA: Harvard University Press, 96–116.
Rawls, John. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Rawls, John. [1993]2005. *Political Liberalism*. New York: Columbia University Press.
Rawls, John. [1997]1999. "The Idea of Public Reason Revisited." In *Collected Papers*. Cambridge, MA: Harvard University Press, 573–615.
Rawls, John. 1999. *The Law of Peoples*. Cambridge, MA: Harvard University Press.
Rawls, John. 2001. *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University Press.
Reidy, David A. 2000. "Rawls's Wide View of Public Reason: Not Wide Enough." *Res Publica* 6: 49–72.
Scanlon, Thomas. 2002. "Rawls on Justification." In *The Cambridge Companion to Rawls*. ed. Samuel Freeman. Cambridge: Cambridge University Press, 139–67. .
Thrasher, John, and Kevin Vallier. 2015. "The Fragility of Consensus: Public Reason, Diversity, and Stability." *European Journal of Philosophy* 23: 933–54.
Vallier, Kevin. 2011. "Convergence and Consensus in Public Reason." *Public Affairs Quarterly* 25: 261–79.
Vallier, Kevin. 2014. *Liberal Politics and Public Faith: Beyond Separation*. New York: Routledge.
Weithman, Paul. 2010. *Why Political Liberalism? On John Rawls's Political Turn*. Oxford: Oxford University Press.
Weithman, Paul. 2015. "Inclusivism, Stability, and Assurance." In *Rawls and Religion*, eds. Tom Bailey and Valentina Gentile. New York: Columbia University Press, 75–98.